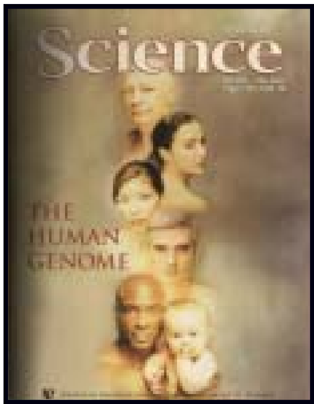


Phylogenomic analysis of disease-resistance proteins in plants

Kimmen Sjölander
University of California Berkeley
http://phylogenomics.berkeley.edu

Collaborators:
David Konerding
Wayne Christopher
Bob Edgar
Joseph Dale
Austin Huang

Collaborators:
Brian Staskawicz
Barbara Baker
Richard Michelmore

Challenges in protein classification

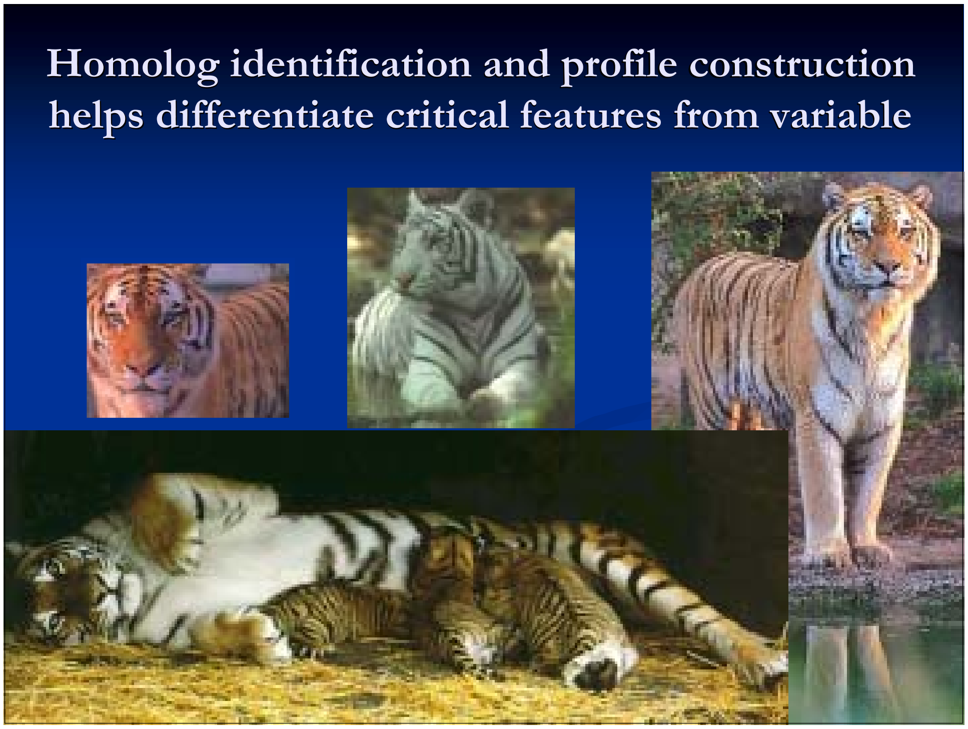
1. Remote homolog detection.
   How much information does knowing a remote homolog provide?
2. Phylogenetic context is critical.
   Paralogs can have divergent function (so can orthologs…)
3. Domain structure issues.
4. Some fraction of the annotations in the sequence databases are not exactly accurate.

Function and Structure Prediction by Homology

If you have a sequence you know nothing about, find a relative.

Given one member, find the relatives...

Would we recognize this member?

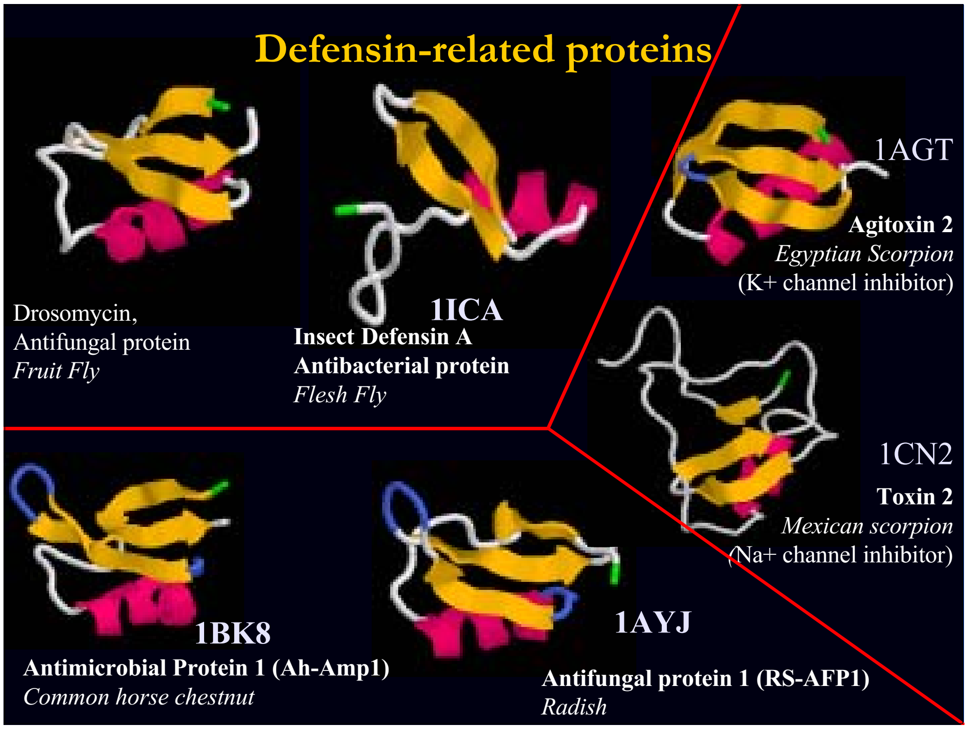


Homolog identification and profile construction helps differentiate critical features from variable

Profile generalization allows us to identify some truly remote relatives

"Evolution conserves structure and function"

But not completely.



Defensin-related proteins

Drosomycin,
Antifungal protein
Fruit Fly

1ICA
Insect Defensin A
Antibacterial protein
Flesh Fly

1AGT
Agitoxin 2
Egyptian Scorpion
(K+ channel inhibitor)

1CN2
Toxin 2
Mexican scorpion
(Na+ channel inhibitor)

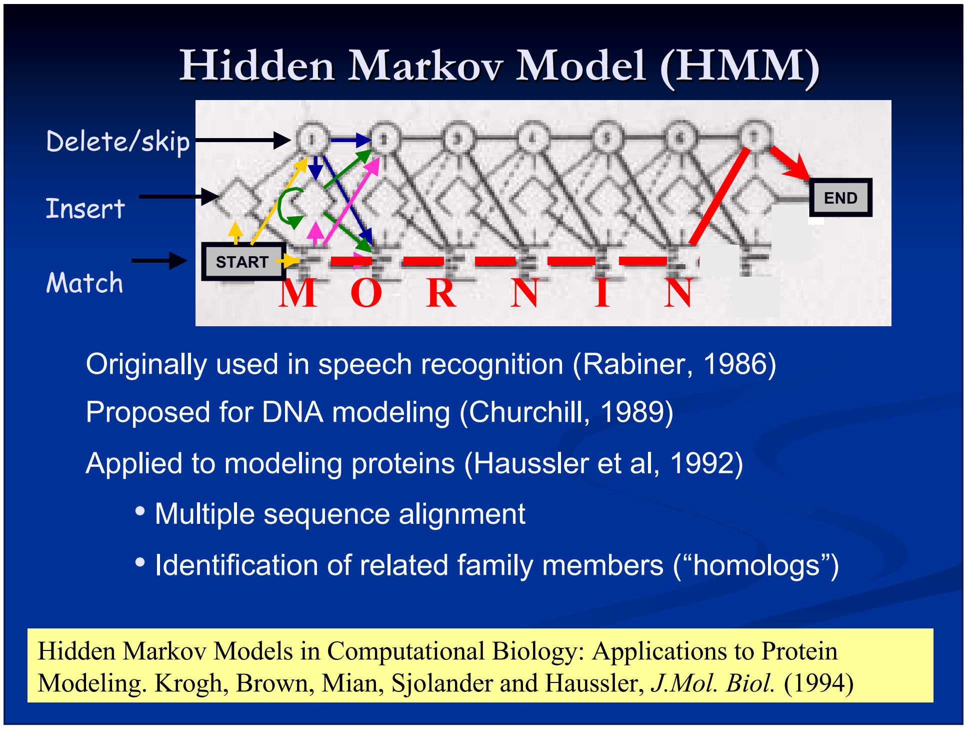
1BK8
Antimicrobial Protein 1 (Ah-Amp1)
Common horse chestnut

1AYJ
Antifungal protein 1 (RS-AFP1)
Radish

Why not just use BLAST?

Poor performance in remote homolog detection compared to HMM methods.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J. Mol. Biol. 284, 1201-1210 (1998).



Hidden Markov Model (HMM)

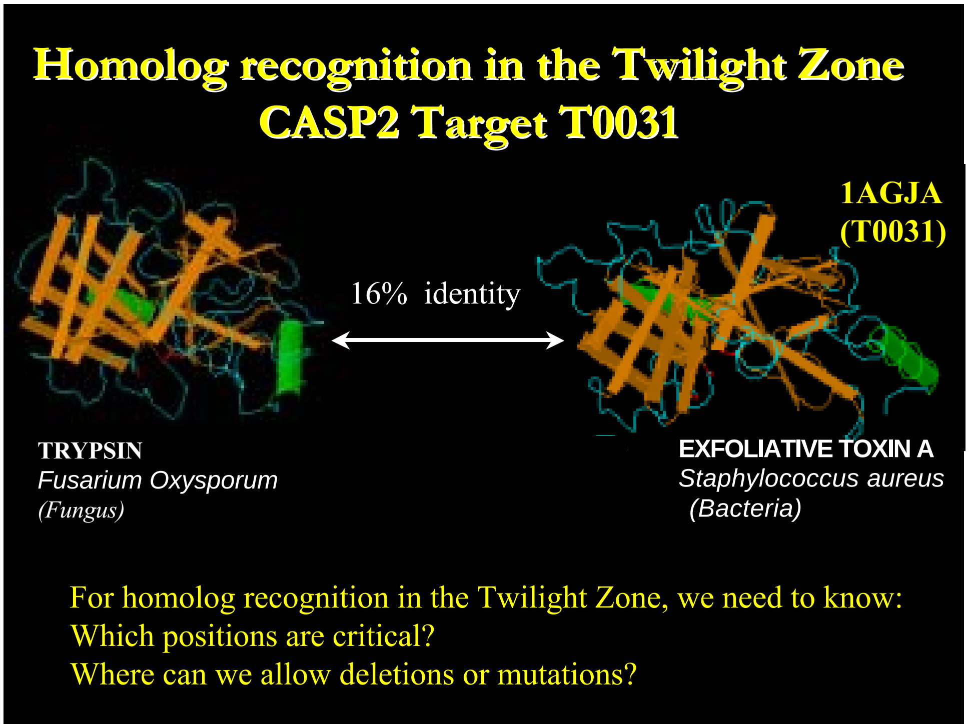
Originally used in speech recognition (Rabiner, 1986)

Proposed for DNA modeling (Churchill, 1989)

Applied to modeling proteins (Haussler et al, 1992)

- Multiple sequence alignment
- Identification of related family members (“homologs”)

Hidden Markov Models in Computational Biology: Applications to Protein Modeling. Krogh, Brown, Mian, Sjolander and Haussler, *J.Mol. Biol.* (1994)



Homolog recognition in the Twilight Zone CASP2 Target T0031

TRYPSIN
Fusarium Oxysporum
(Fungus)

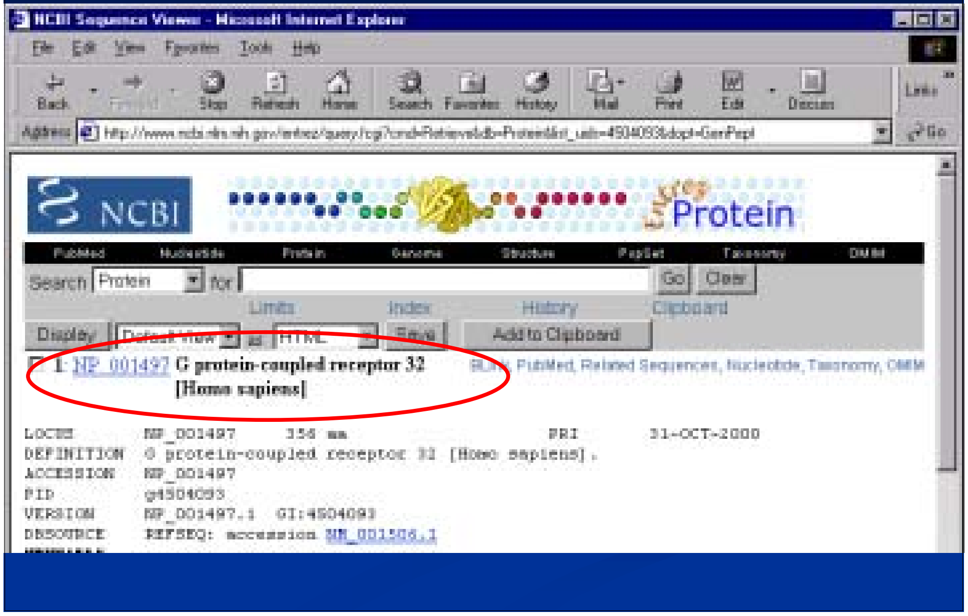
EXFOLIATIVE TOXIN A
Staphylococcus aureus
(Bacteria)

For homolog recognition in the Twilight Zone, we need to know:
Which positions are critical?
Where can we allow deletions or mutations?

Homolog detection is just the first step…

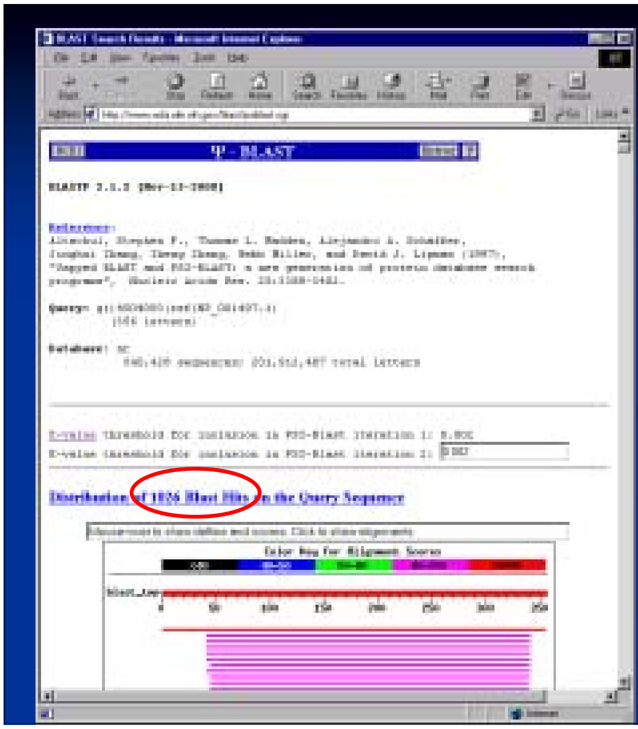
Correct functional classification requires attention to evolutionary relationships

Example 1: Orphan GPCR classification


NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Links
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=4504093&dopt=GenPept
Go
NCBI
Protein
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Search Protein for
Go Clear
Limits Index History Clipboard
Display Default View as HTML Save Add to Clipboard
BLink, PubMed, Related Sequences, Nucleotide, Taxonomy, OMIM

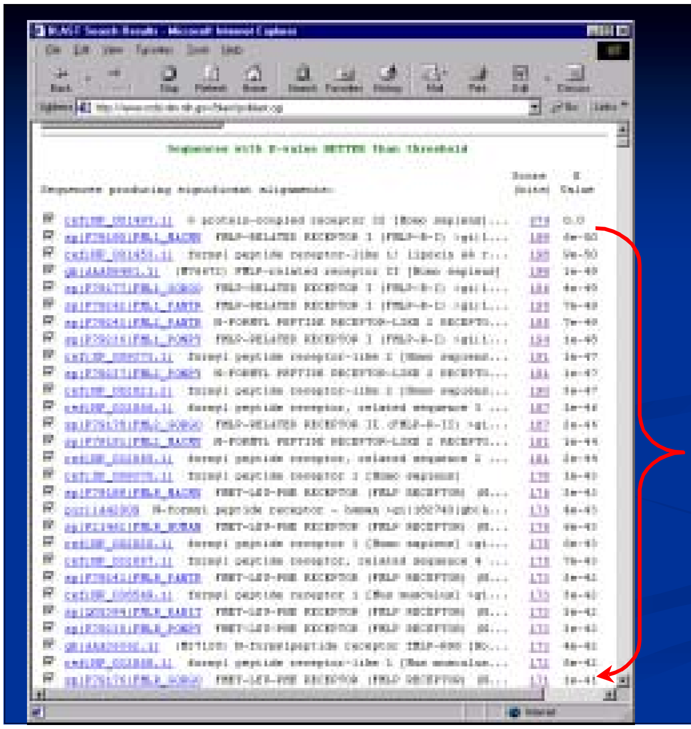
1: NP_001497 G protein-coupled receptor 32 [Homo sapiens]

```
LOCUS       NP_001497     356 aa                    PRI       31-OCT-2000
DEFINITION  G protein-coupled receptor 32 [Homo sapiens].
ACCESSION   NP_001497
PID         g4504093
VERSION     NP_001497.1  GI:4504093
DBSOURCE    REFSEQ: accession NM_001506.1
```



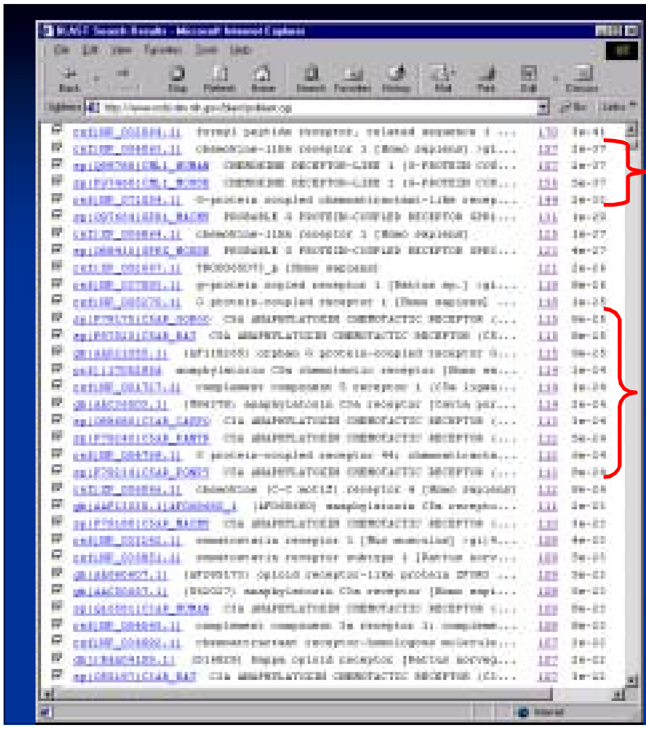
```
Ψ - BLAST

BLASTP 2.1.2 [Nov-13-2000]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= gi|4503783|ref|NP_001727.1|
         (350 letters)

Database: nr
           640,425 sequences; 201,512,487 total letters

E-value threshold for inclusion in PSI-Blast iteration 1: 0.005
E-value threshold for inclusion in PSI-Blast iteration 2: 0.005

Distribution of 1026 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments
```

Step 1:
Run
BLAST



```
                    Sequences with E-value BETTER than threshold

                                                                    Score     E
Sequences producing significant alignments:                        (bits)  Value

ref|NP_001727.1|  G protein-coupled receptor 32 [Homo sapiens]...   [illegible]  0.0
sp|[illegible]|FML1_MACMU  FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi|...   [illegible]  4e-50
ref|NP_[illegible]|  formyl peptide receptor-like 1; lipoxin A4 r...   [illegible]  9e-50
gb|AAA[illegible]|  (M76672) FMLP-related receptor II [Homo sapiens]   [illegible]  1e-49
sp|[illegible]|FML1_GORGO  FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi|...   [illegible]  4e-49
sp|[illegible]|FML1_PANTR  FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi|...   [illegible]  7e-49
sp|[illegible]|FML2_PANTR  N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...   [illegible]  7e-49
sp|[illegible]|FML1_PONPY  FMLP-RELATED RECEPTOR I (FMLP-R-I) >gi|...   [illegible]  1e-48
ref|NP_[illegible]|  formyl peptide receptor-like 2 [Homo sapiens]...   [illegible]  1e-47
sp|[illegible]|FML2_PONPY  N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...   [illegible]  1e-47
ref|NP_[illegible]|  formyl peptide receptor-like 2 [Homo sapiens]...   [illegible]  5e-47
ref|NP_[illegible]|  formyl peptide receptor, related sequence 1 ...   [illegible]  1e-46
sp|[illegible]|FML2_GORGO  FMLP-RELATED RECEPTOR II (FMLP-R-II) >gi|...   [illegible]  2e-45
sp|[illegible]|FML2_MACMU  N-FORMYL PEPTIDE RECEPTOR-LIKE 2 RECEPTO...   [illegible]  1e-44
ref|NP_[illegible]|  formyl peptide receptor, related sequence 2 ...   [illegible]  2e-44
ref|NP_[illegible]|  formyl peptide receptor 1 [Homo sapiens]   [illegible]  1e-43
sp|[illegible]|FMLR_MACMU  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  1e-43
pir||[illegible]  N-formyl peptide receptor - human >gi|[illegible]|gb|A...   [illegible]  4e-43
sp|[illegible]|FMLR_HUMAN  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  4e-43
ref|NP_[illegible]|  formyl peptide receptor 1 [Homo sapiens] >gi...   [illegible]  4e-43
ref|NP_[illegible]|  formyl peptide receptor, related sequence 4 ...   [illegible]  7e-43
sp|[illegible]|FMLR_PANTR  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  1e-42
ref|NP_[illegible]|  formyl peptide receptor 1 [Mus musculus] >gi|...   [illegible]  1e-42
sp|[illegible]|FMLR_RABIT  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  1e-42
sp|[illegible]|FMLR_PONPY  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  1e-42
gb|AAA[illegible]|  (M71109) N-formylpeptide receptor IMLP-R98 [Ho...   [illegible]  4e-42
ref|NP_[illegible]|  formyl peptide receptor-like 1 [Mus musculus...   [illegible]  4e-42
sp|[illegible]|FMLR_GORGO  FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR) (N...   [illegible]  1e-41
```

Is the
query an
FMLP
receptor?

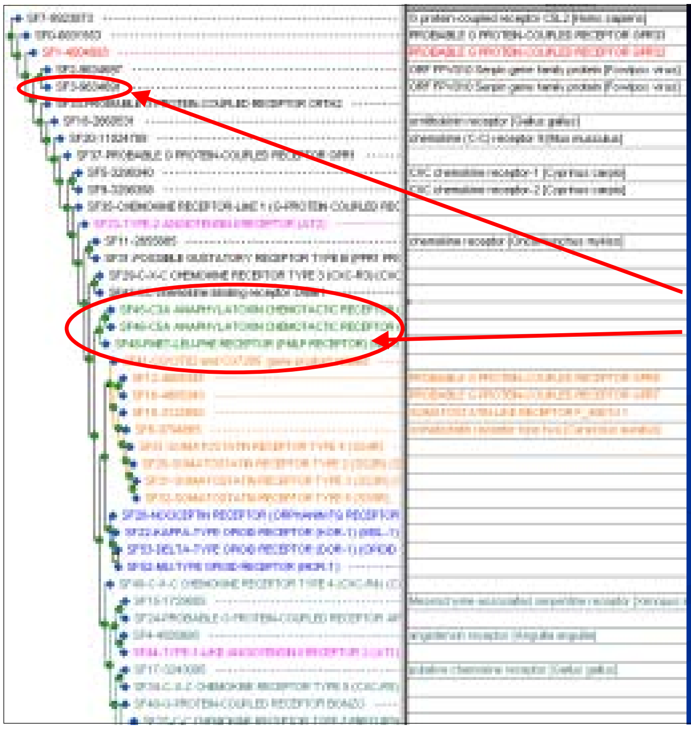


```
[illegible]  formyl peptide receptor, related sequence 1 ...
[illegible]  chemokine-like receptor 1 [Homo sapiens] >gi...
[illegible]  CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COU...
[illegible]  CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COU...
[illegible]  G-protein coupled chemoattractant-like recep...
[illegible]  PROBABLE G PROTEIN-COUPLED RECEPTOR GPR...
[illegible]  chemokine-like receptor 1 [Homo sapiens]
[illegible]  PROBABLE G PROTEIN-COUPLED RECEPTOR GPR...
[illegible]  [illegible] [Homo sapiens]
[illegible]  g-protein coupled receptor 1 [Rattus sp.] >gi...
[illegible]  G protein-coupled receptor 1 [Homo sapiens] ...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C5...
[illegible]  [illegible] orphan G protein-coupled receptor G...
[illegible]  anaphylatoxin C5a chemotactic receptor [Homo sa...
[illegible]  complement component 5 receptor 1 (C5a ligan...
[illegible]  anaphylatoxin C5a receptor [Cavia por...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  G protein-coupled receptor 44; chemoattracta...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  chemokine (C-C motif) receptor 4 [Homo sapiens]
[illegible]  [illegible] anaphylatoxin C5a recepto...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  somatostatin receptor 5 [Mus musculus] >gi|...
[illegible]  somatostatin receptor subtype 1 [Rattus norv...
[illegible]  opioid receptor-like protein ...
[illegible]  anaphylatoxin C5a receptor [Homo sapi...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (...
[illegible]  complement component 3a receptor 1; complem...
[illegible]  chemoattractant receptor-homologous molecule...
[illegible]  kappa opioid receptor [Rattus norveg...
[illegible]  C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C5...
```

Or a chemokine receptor?

Or a C5A Anaphylatoxin chemotactic receptor?

Or…?



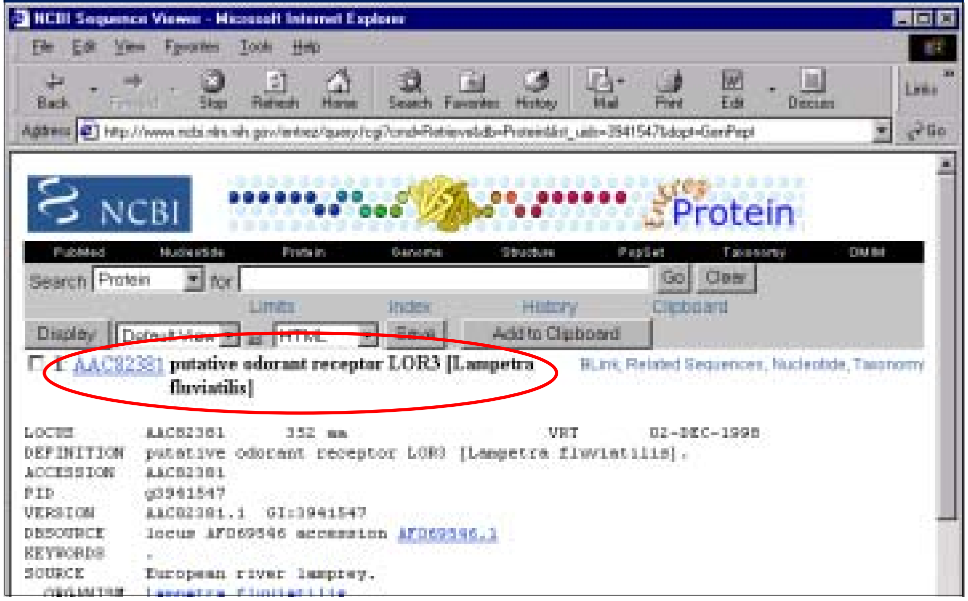
PROBABLE G PROTEIN-COUPLED RECEPTOR
CHEMOKINE RECEPTOR-LIKE 1 (G-PROTEIN COUPLED REC
C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR
C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR
FMET-LEU-PHE RECEPTOR (FMLP RECEPTOR)
SOMATOSTATIN RECEPTOR TYPE
NOCICEPTIN RECEPTOR (ORPHANIN FQ RECEPTOR
KAPPA-TYPE OPIOID RECEPTOR
DELTA-TYPE OPIOID RECEPTOR
MU-TYPE OPIOID RECEPTOR
C-X-C CHEMOKINE RECEPTOR TYPE 4
G protein-coupled receptor
chemokine receptor

Or a novel type of receptor?

Another reason to not rely on pairwise sequence similarity

What if the top-scoring match is incorrectly annotated?

Example 2: Errors in database annotations

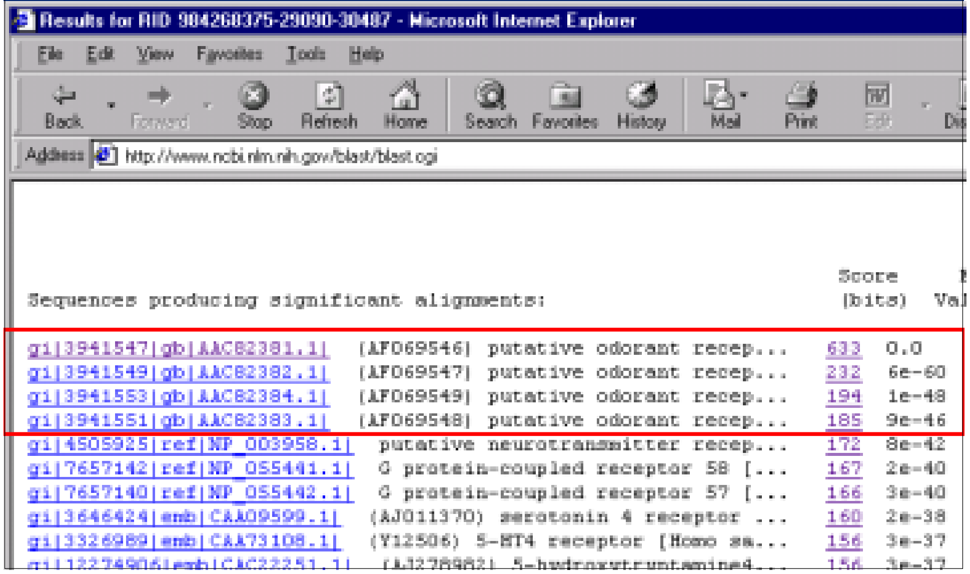


NCBI Sequence Viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&list_uids=3941547&dopt=GenPept
Go Links
NCBI
Protein
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Search Protein for Go Clear
Limits Index History Clipboard
Display Default View HTML Save Add to Clipboard
BLink, Related Sequences, Nucleotide, Taxonomy

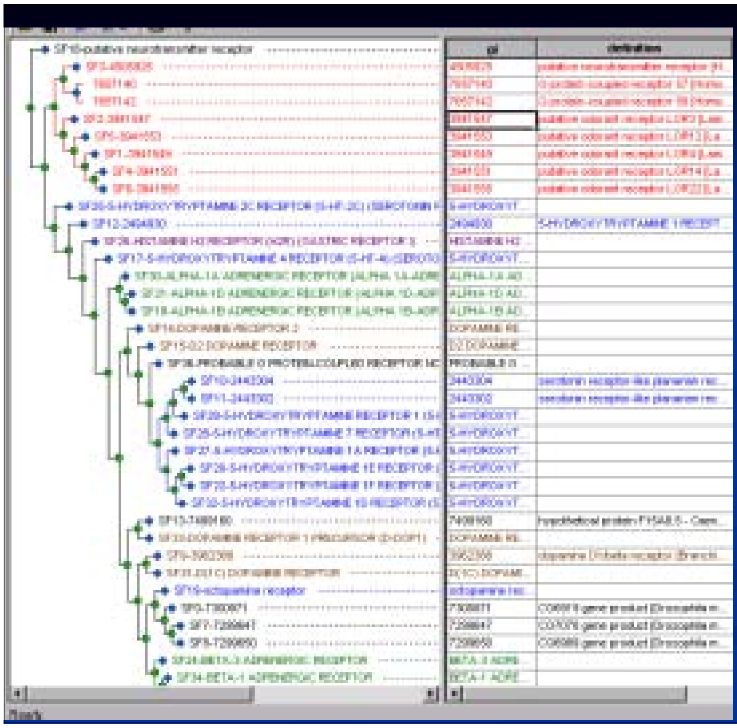
□ 1: AAC82381 putative odorant receptor LOR3 [Lampetra fluviatilis]

```
LOCUS       AAC82381      352 aa                    VRT       02-DEC-1998
DEFINITION  putative odorant receptor LOR3 [Lampetra fluviatilis].
ACCESSION   AAC82381
PID         g3941547
VERSION     AAC82381.1  GI:3941547
DBSOURCE    locus AF069546 accession AF069546.1
KEYWORDS    .
SOURCE      European river lamprey.
  ORGANISM  Lampetra fluviatilis
```

The top matching BLAST hits are also putative odorant receptors

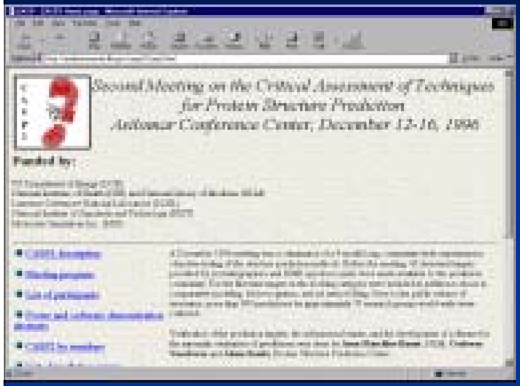

File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Edit
Di
Address
http://www.ncbi.nlm.nih.gov/blast/blast.cgi

```
Results for RID 984268375-29090-30487 - Microsoft Internet Explorer

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Val

gi|3941547|gb|AAC82381.1|   (AF069546) putative odorant recep...    633   0.0
gi|3941549|gb|AAC82382.1|   (AF069547) putative odorant recep...    232   6e-60
gi|3941553|gb|AAC82384.1|   (AF069549) putative odorant recep...    194   1e-48
gi|3941551|gb|AAC82383.1|   (AF069548) putative odorant recep...    185   9e-46
gi|4505925|ref|NP_003958.1|  putative neurotransmitter recep...     172   8e-42
gi|7657142|ref|NP_055441.1|  G protein-coupled receptor 58 [...     167   2e-40
gi|7657140|ref|NP_055442.1|  G protein-coupled receptor 57 [...     166   3e-40
gi|3646424|emb|CAA09599.1|  (AJ011370) serotonin 4 receptor ...     160   2e-38
gi|3326989|emb|CAA73108.1|  (Y12506) 5-HT4 receptor [Homo sa...     156   3e-37
gi|12274906|emb|CAC22251.1|  (AJ278982) 5-hydroxytryptamine4...     156   3e-37
```


gi
definition
SF10-putative neurotransmitter receptor
7657140
7657142
SF12-2494830
SF14-DOPAMINE RECEPTOR 2
SF15-D2 DOPAMINE RECEPTOR
SF19-octopamine receptor
SF24-BETA-3 ADRENERGIC RECEPTOR
SF34-BETA-1 ADRENERGIC RECEPTOR
5-HYDROXYT...
ALPHA-1A AD...
DOPAMINE RE...
PROBABLE G...
putative odorant receptor
5-HYDROXYTRYPTAMINE 1 RECEPT
hypothetical protein F16A11.5 - Caen...
dopamine D1/beta receptor (Branchi...
Ready

Phylogenetic analysis suggests it's more likely a Biogenic Amine GPCR

Lessons from CASP2



Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction
Asilomar Conference Center, December 12-16, 1996

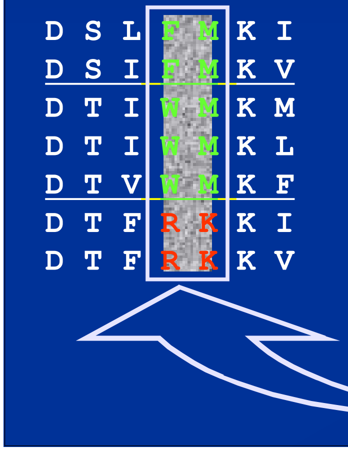
Funded by:

[illegible]

1. HMMs optimized for remote homolog detection generally require clustering and alignment of many divergent sequences.
2. Alignments of new sequences to these HMMs can be pretty awful.

Given a protein sequence ("target"), predict its most likely fold, and produce an alignment of the target and the solved structure. Predictions judged using structure-structure alignment (SCOP, VAST, DALI).

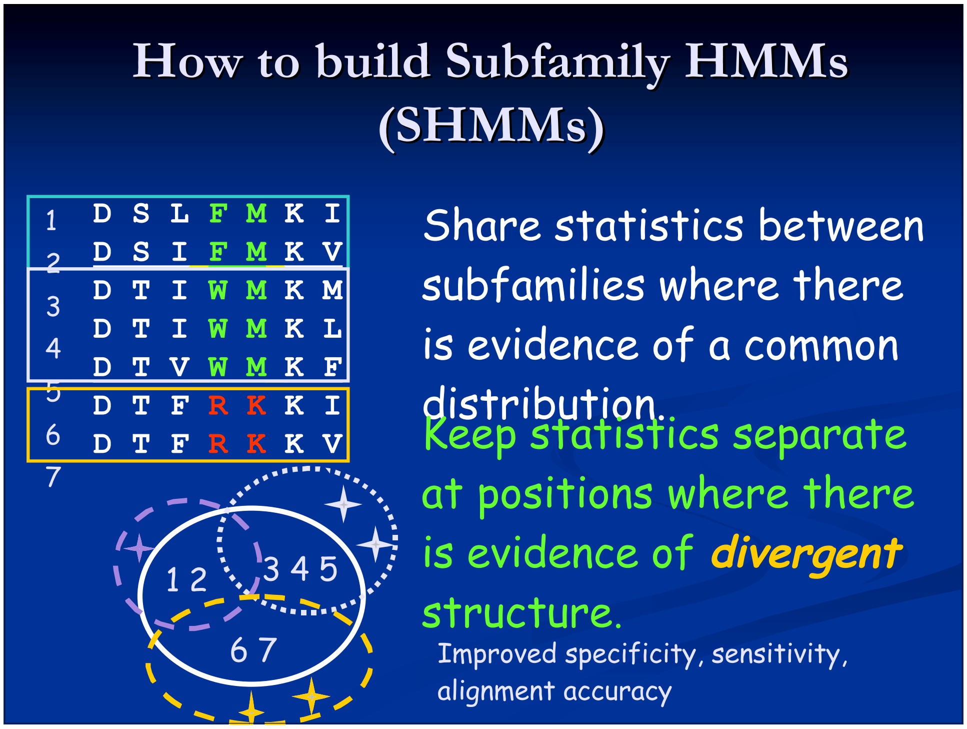
Conflict



```
D S L F M K I
D S I F M K V
D T I W M K M
D T I W M K L
D T V W M K F
D T F R K K I
D T F R K K V
```

- For effective *remote homolog* detection, a profile or HMM needs information from divergent family members
- Without this context, we cannot differentiate critical from variable positions
- HMMs constructed with such data provide a coarse classification
- But, the more variability we introduce in training data, the greater the potential ***noise*** at some positions

Subfamily HMM construction



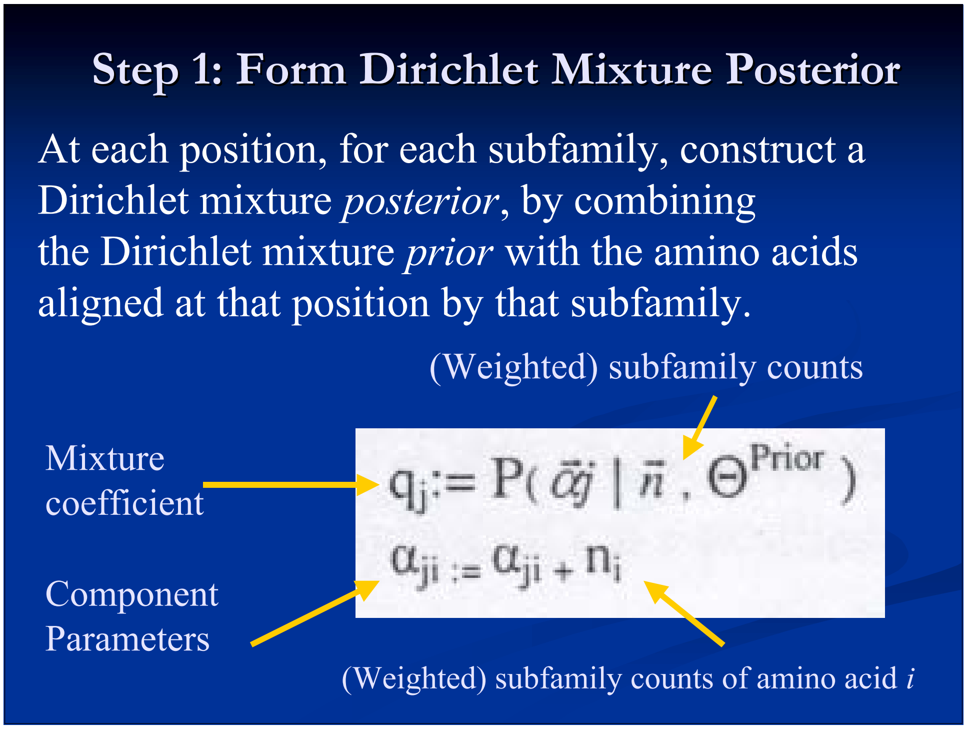
How to build Subfamily HMMs (SHMMs)

```
1  D S L F M K I
2  D S I F M K V
3  D T I W M K M
4  D T I W M K L
5  D T V W M K F
6  D T F R K K I
7  D T F R K K V
```

Share statistics between subfamilies where there is evidence of a common distribution.

Keep statistics separate at positions where there is evidence of ***divergent*** structure.

Improved specificity, sensitivity, alignment accuracy



Step 1: Form Dirichlet Mixture Posterior

At each position, for each subfamily, construct a Dirichlet mixture *posterior*, by combining the Dirichlet mixture *prior* with the amino acids aligned at that position by that subfamily.

(Weighted) subfamily counts

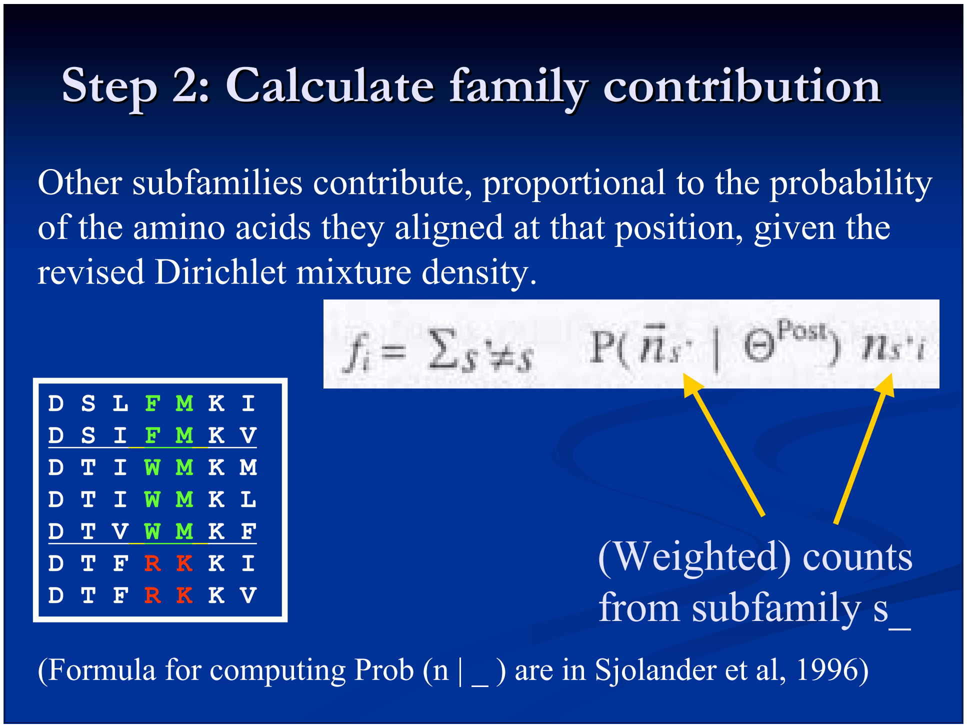
Mixture coefficient

$$q_j := P(\vec{\alpha}_j \mid \vec{n}, \Theta^{\text{Prior}})$$

$$\alpha_{ji} := \alpha_{ji} + n_i$$

Component Parameters

(Weighted) subfamily counts of amino acid *i*



Step 2: Calculate family contribution

Other subfamilies contribute, proportional to the probability of the amino acids they aligned at that position, given the revised Dirichlet mixture density.

$$f_i = \Sigma_{s' \neq s} \quad P(\vec{n}_{s'} \mid \Theta^{\text{Post}})\ n_{s'i}$$

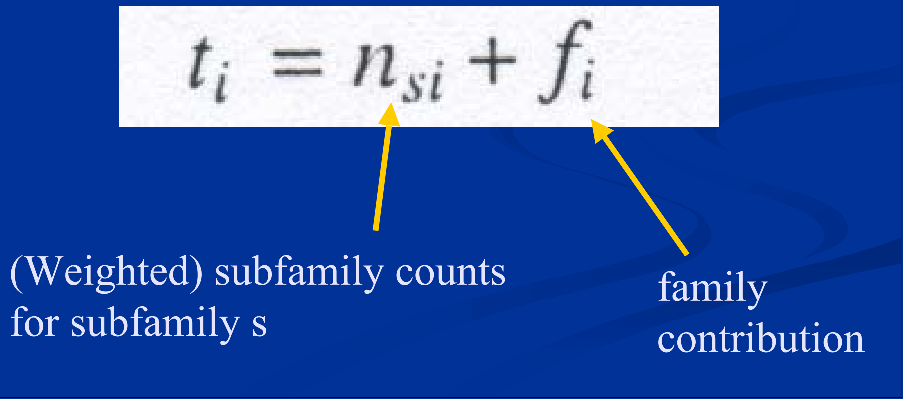
```
D S L F M K I
D S I F M K V
D T I W M K M
D T I W M K L
D T V W M K F
D T F R K K I
D T F R K K V
```

(Weighted) counts from subfamily s_

(Formula for computing Prob (n | _) are in Sjolander et al, 1996)

Step 3: Compute pseudocounts

Add the family contribution to the observed (weighted) counts, to obtain the pseudocounts t_i of amino acid i:



$$t_i = n_{si} + f_i$$

(Weighted) subfamily counts for subfamily s

family contribution

Step 4: Compute amino acid probabilities

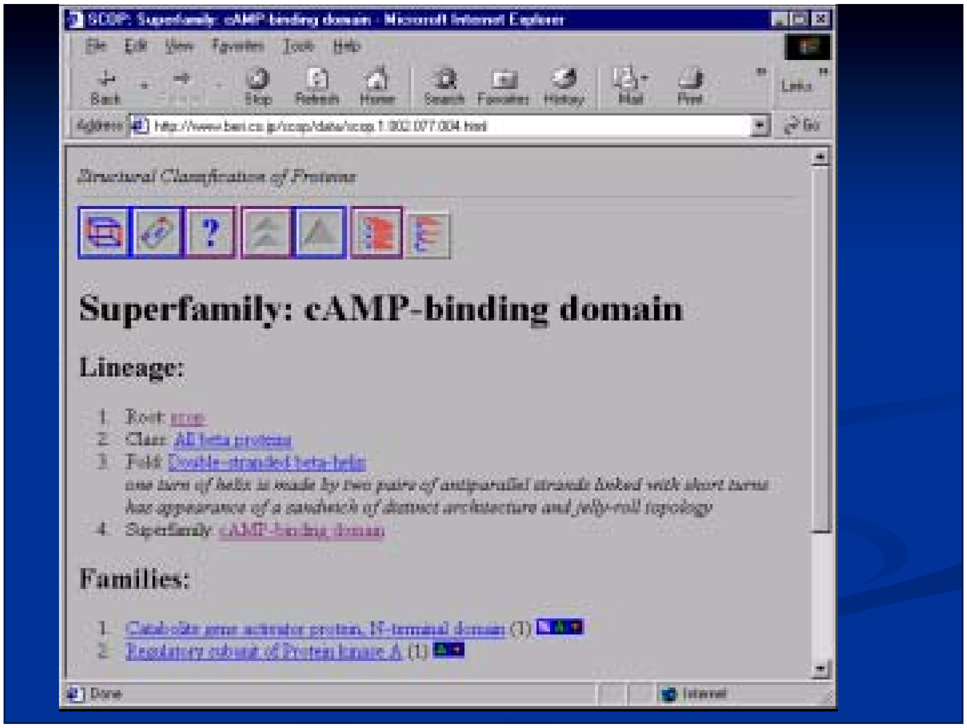
Normally, we compute amino acid probabilities by combining a Dirichlet mixture prior with observed counts as follows:

$$\hat{p}_i = \sum_j P(\vec{\alpha}_j \mid \vec{n}) \frac{n_i + \alpha_{ji}}{|\vec{n}| + |\vec{\alpha}_j|}$$

Instead, we will estimate the probability of amino acid *i* as follows:

$$\hat{p}_i = \sum_j P(\vec{\alpha}_j \mid \vec{n}) \frac{t_i + \alpha_{ji}}{|\vec{t}| + |\vec{\alpha}_j|}$$

Subfamily HMM Performance


SCOP: Superfamily: cAMP-binding domain - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Links
Address
http://www.beri.co.jp/scop/data/scop.1.002.077.004.html
Go
Done
Internet

Structural Classification of Proteins

Superfamily: cAMP-binding domain

Lineage:

1. Root: scop
2. Class: All beta proteins
3. Fold: Double-stranded beta-helix
   one turn of helix is made by two pairs of antiparallel strands linked with short turns
   has appearance of a sandwich of distinct architecture and jelly-roll topology
4. Superfamily: cAMP-binding domain

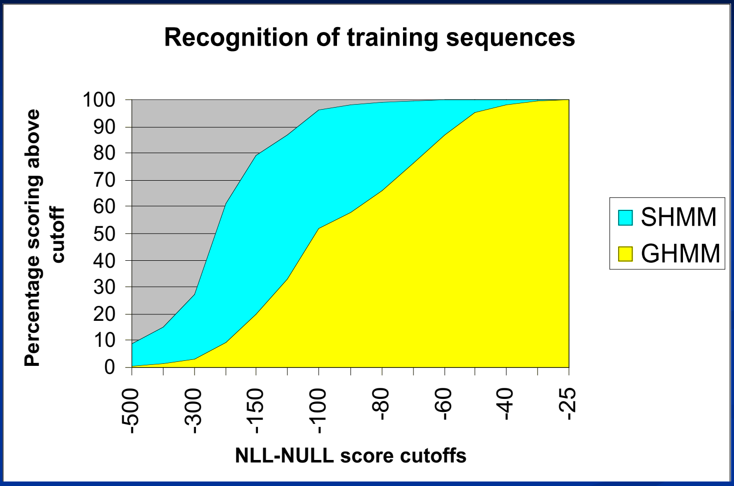
Families:

1. Catabolite gene activator protein, N-terminal domain (1)
2. Regulatory subunit of Protein kinase A (1)

Socrates' First Command: Know Thyself

Test 1. How accurate are subfamily HMMs at recognizing their own training sequences?

Test 1: Training Sequence Recognition



% sequences found above cutoffs: SHMM method 100%
GHMM method 99.89%

Error (in subfamily assignment): 35/9,465 (~3.6/1,000)

Honor thy father and thy mother

and thy brothers and sisters

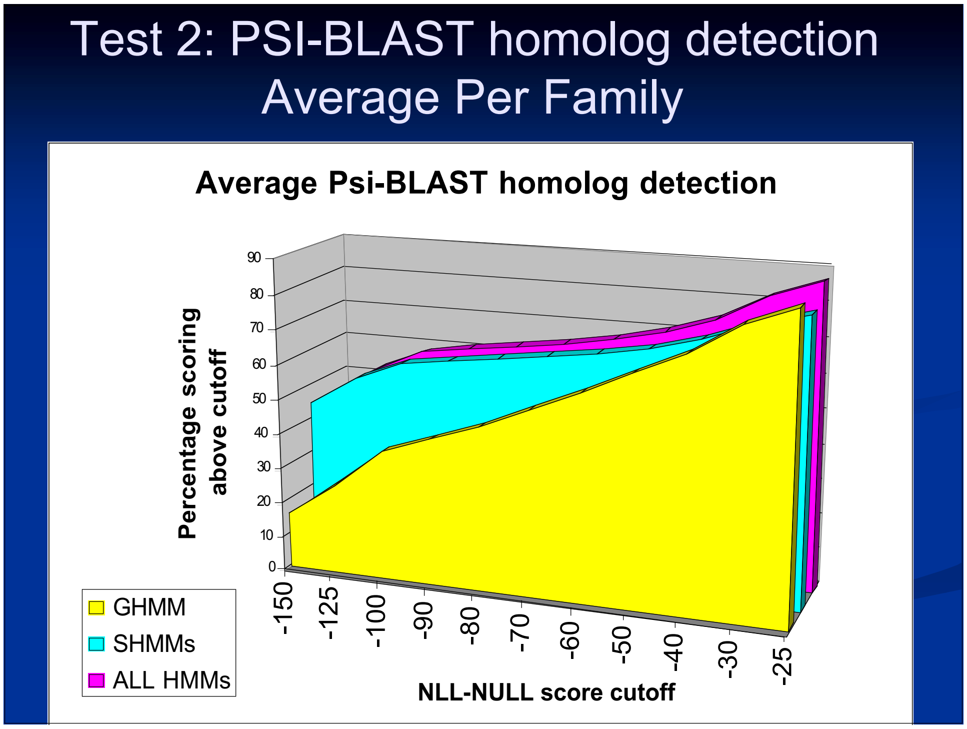
and aunts and uncles

and cousins

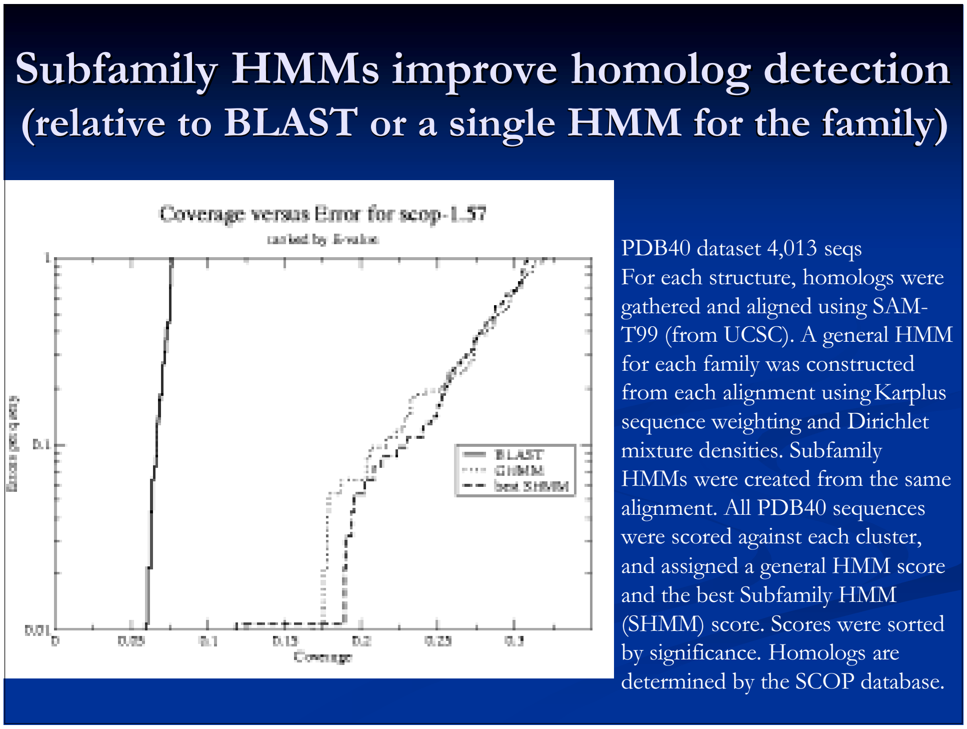
and second cousins

and third cousins twice removed…

Recognition and classification of family members



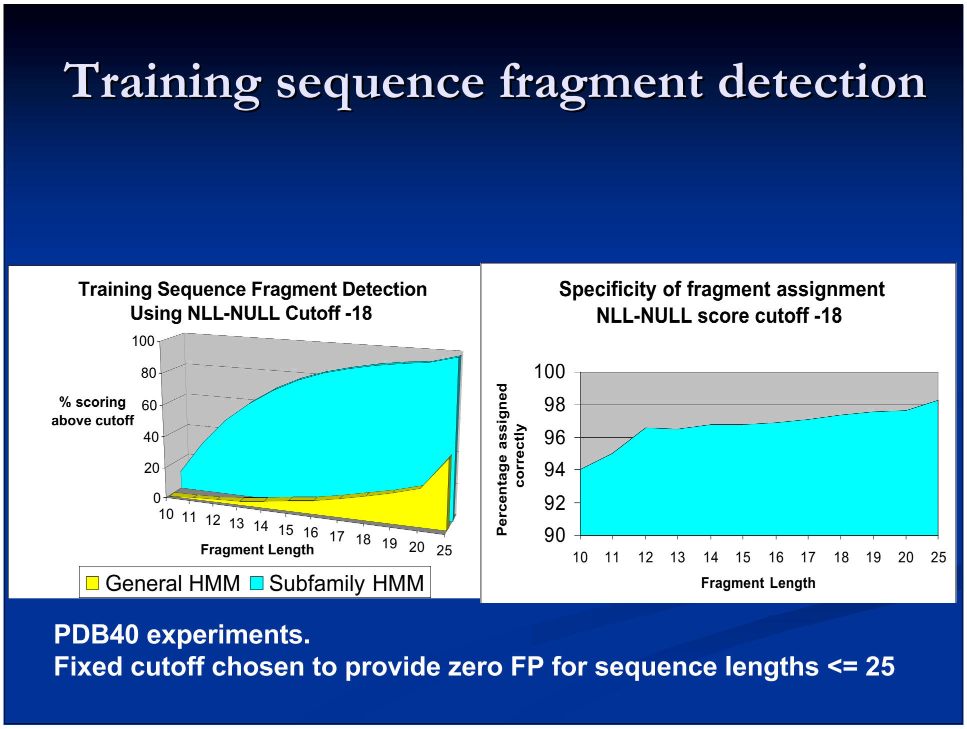
Test 2: PSI-BLAST homolog detection Average Per Family



Subfamily HMMs improve homolog detection (relative to BLAST or a single HMM for the family)

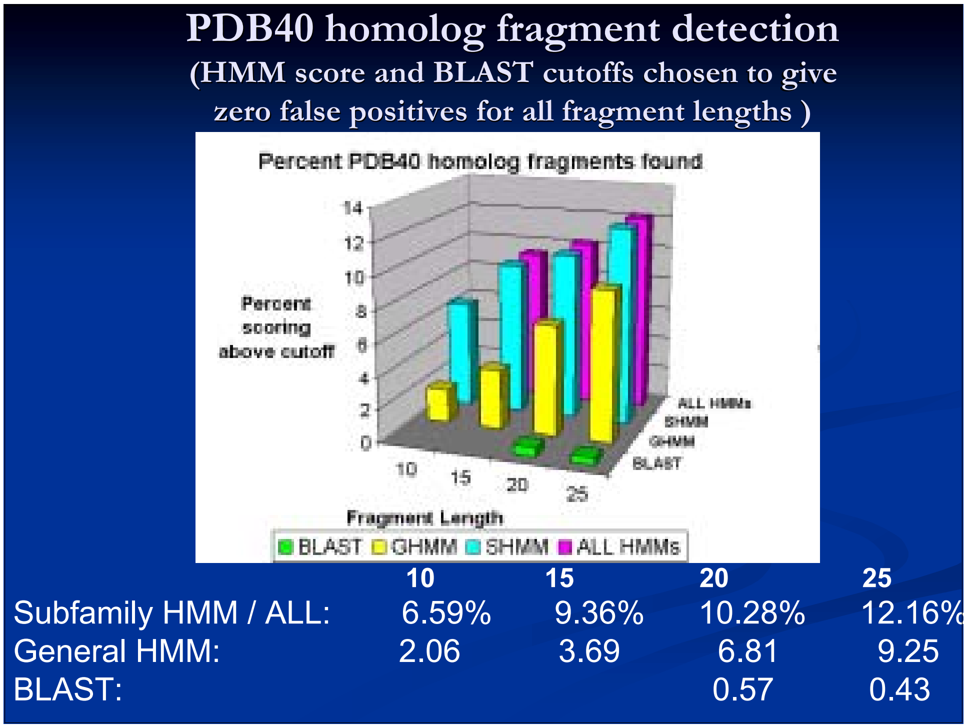
PDB40 dataset 4,013 seqs
For each structure, homologs were gathered and aligned using SAM-T99 (from UCSC). A general HMM for each family was constructed from each alignment using Karplus sequence weighting and Dirichlet mixture densities. Subfamily HMMs were created from the same alignment. All PDB40 sequences were scored against each cluster, and assigned a general HMM score and the best Subfamily HMM (SHMM) score. Scores were sorted by significance. Homologs are determined by the SCOP database.

Fragments and ESTs can be especially challenging



Training sequence fragment detection

PDB40 experiments.
Fixed cutoff chosen to provide zero FP for sequence lengths <= 25



PDB40 homolog fragment detection

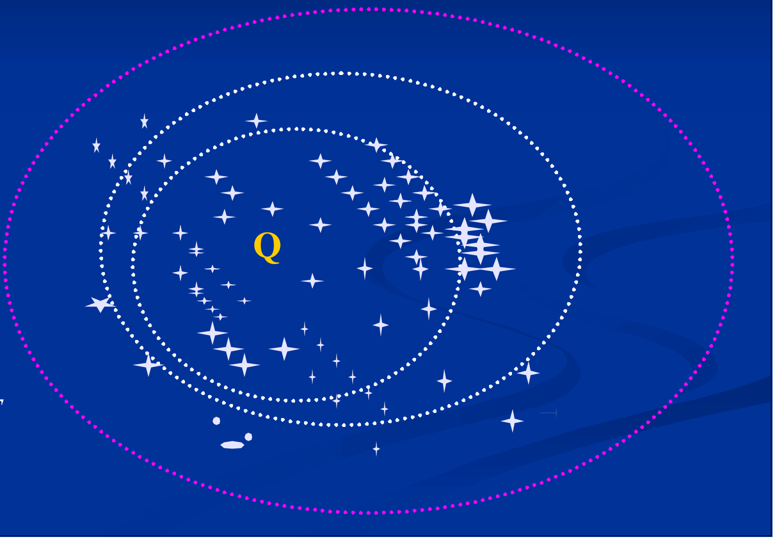
(HMM score and BLAST cutoffs chosen to give zero false positives for all fragment lengths)

	10	15	20	25
Subfamily HMM / ALL:	6.59%	9.36%	10.28%	12.16%
General HMM:	2.06	3.69	6.81	9.25
BLAST:			0.57	0.43

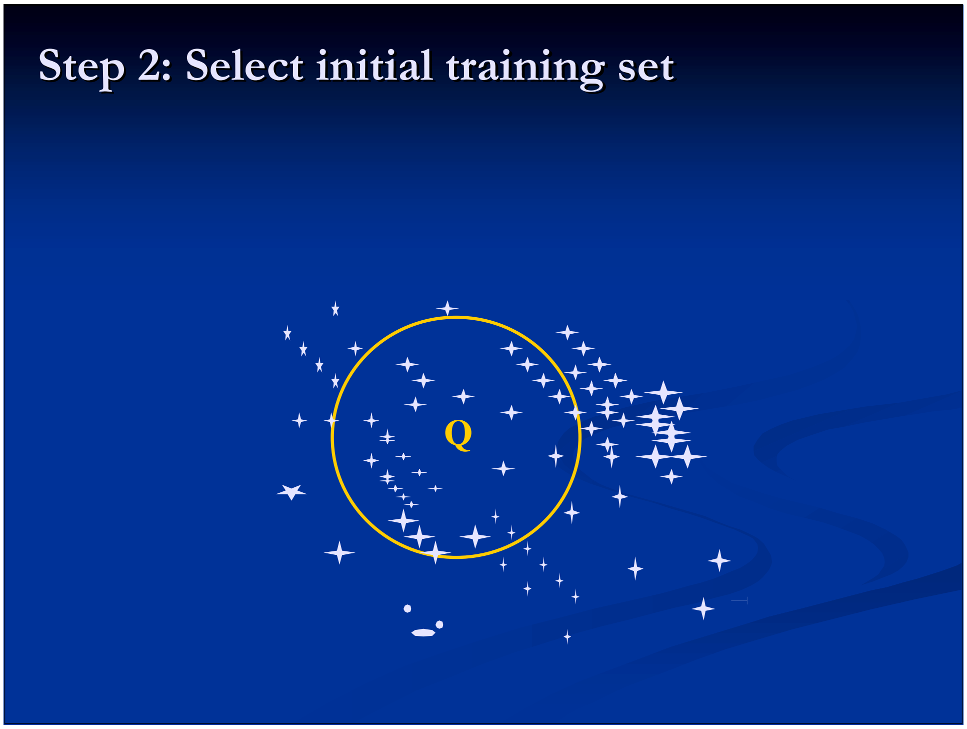
FlowerPower

Iterative clustering and alignment tool

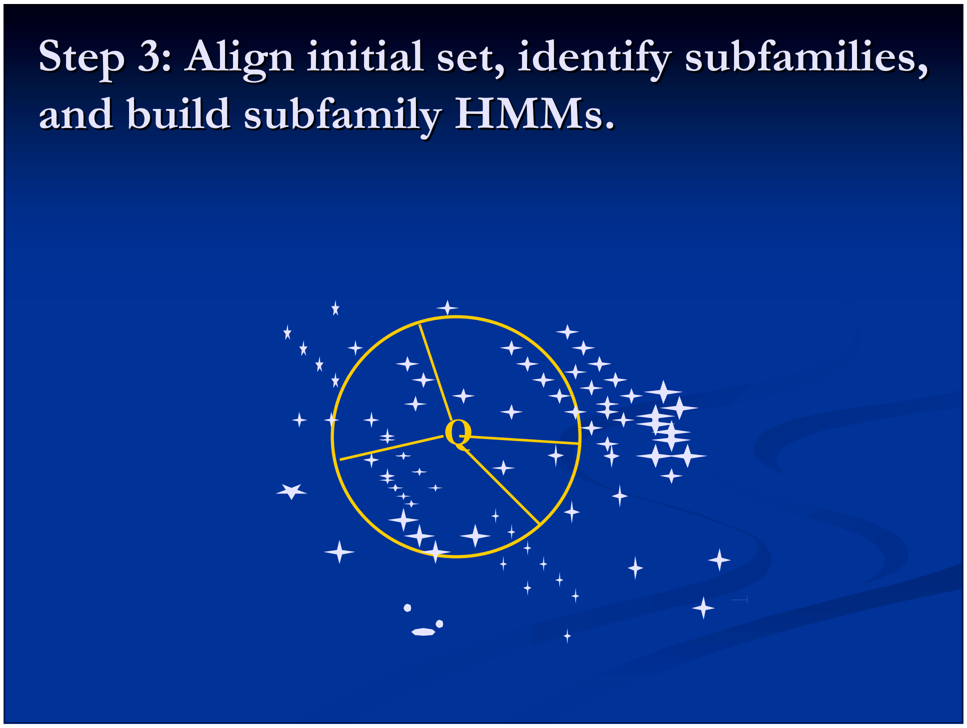
Step 1: Identify putative homologs to query sequence (Q)



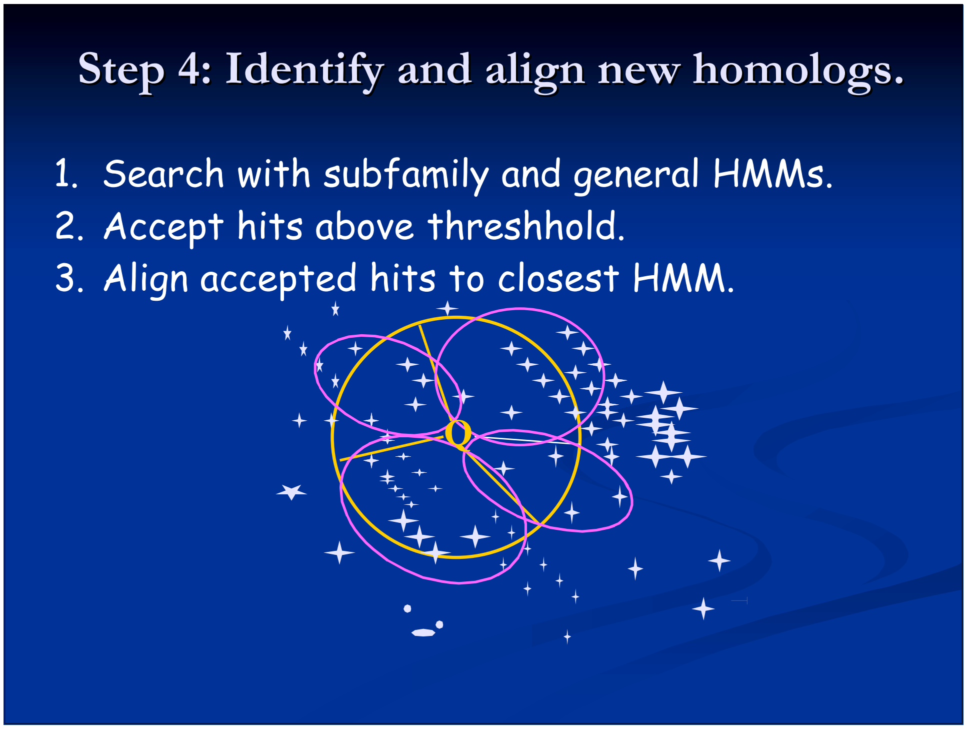
PSI-BLAST
3 iterations
E-5 cutoff



Step 2: Select initial training set

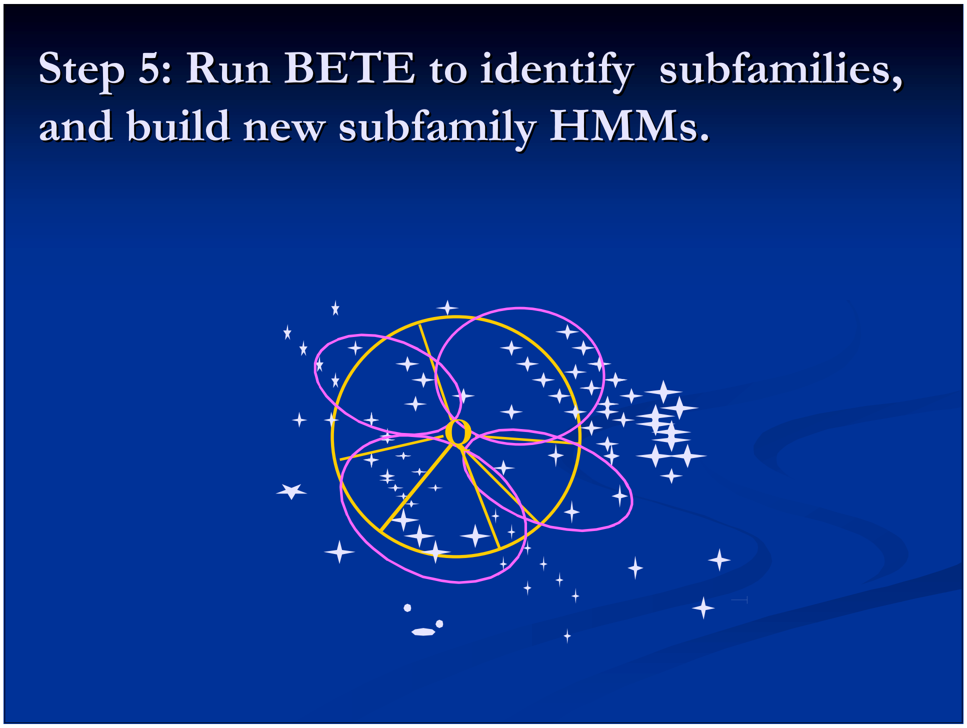


Step 3: Align initial set, identify subfamilies, and build subfamily HMMs.

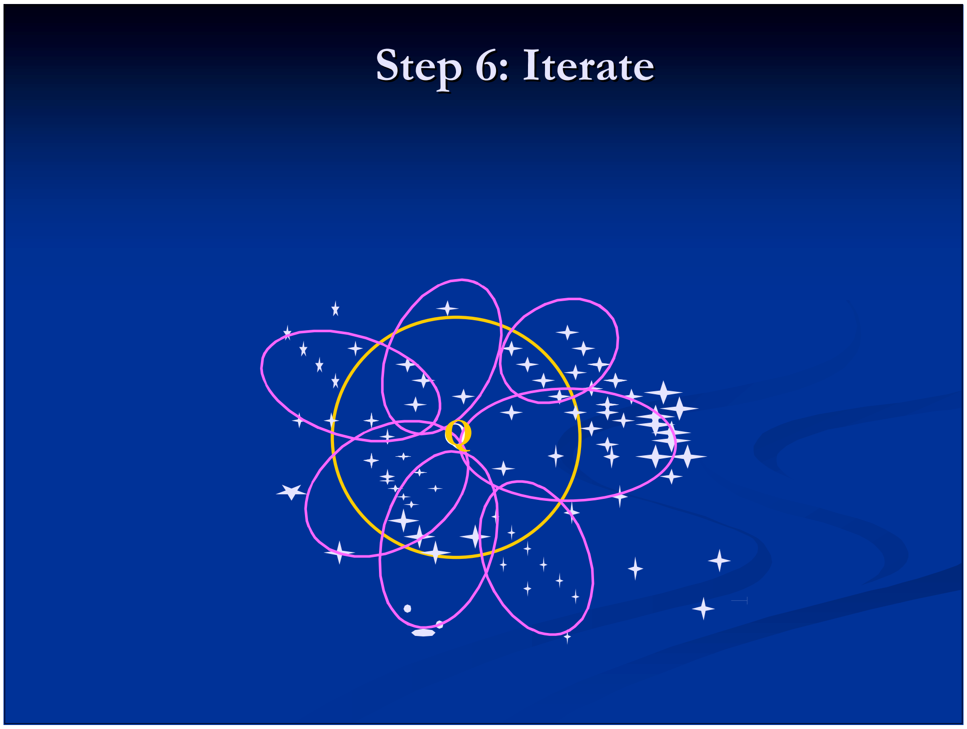


Step 4: Identify and align new homologs.

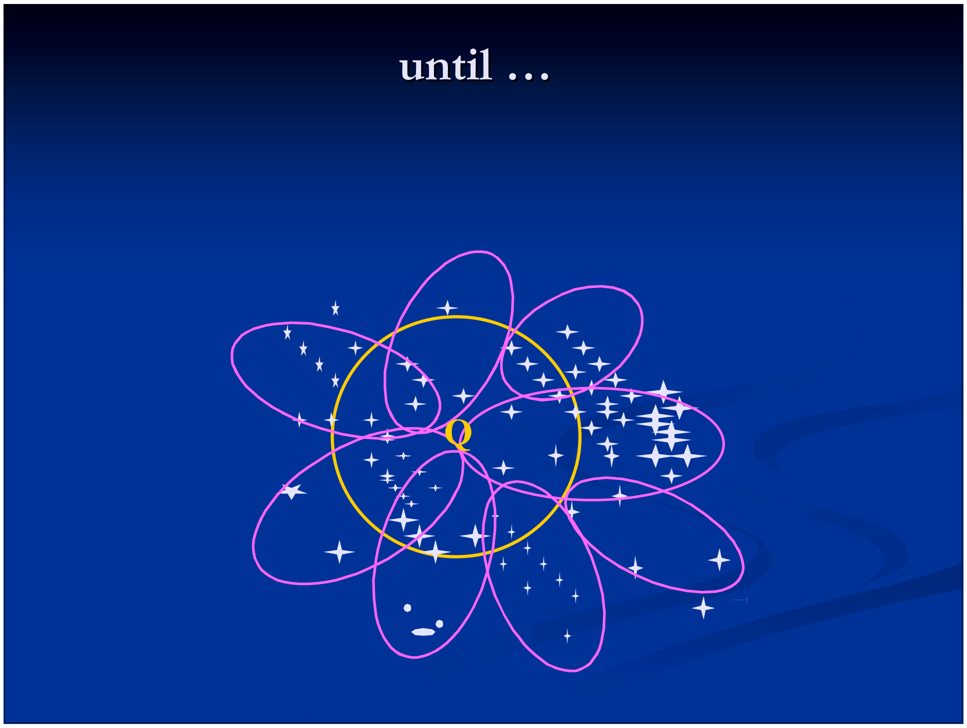
1. Search with subfamily and general HMMs.
2. Accept hits above threshhold.
3. Align accepted hits to closest HMM.



Step 5: Run BETE to identify subfamilies, and build new subfamily HMMs.

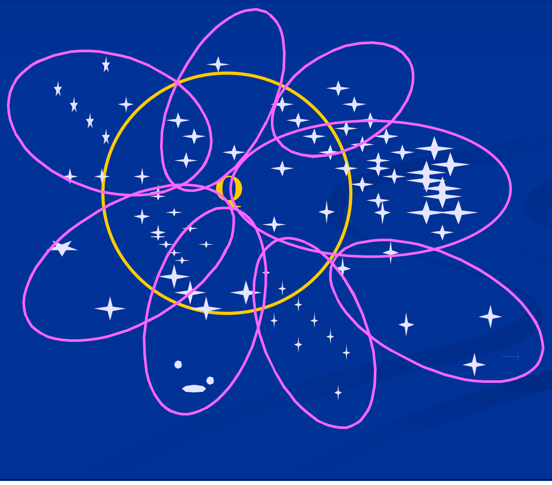


Step 6: Iterate

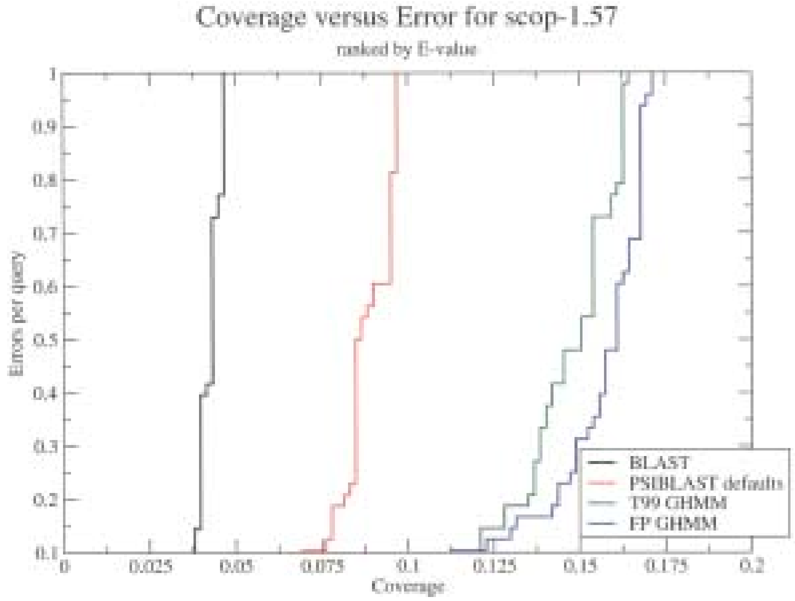


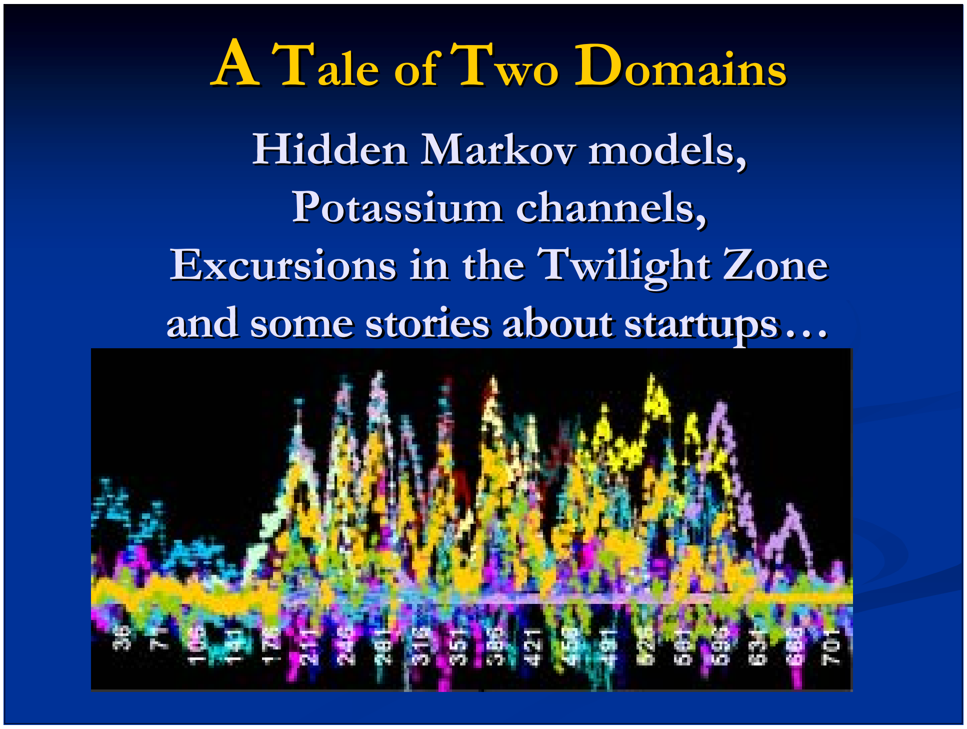
until …

convergence.



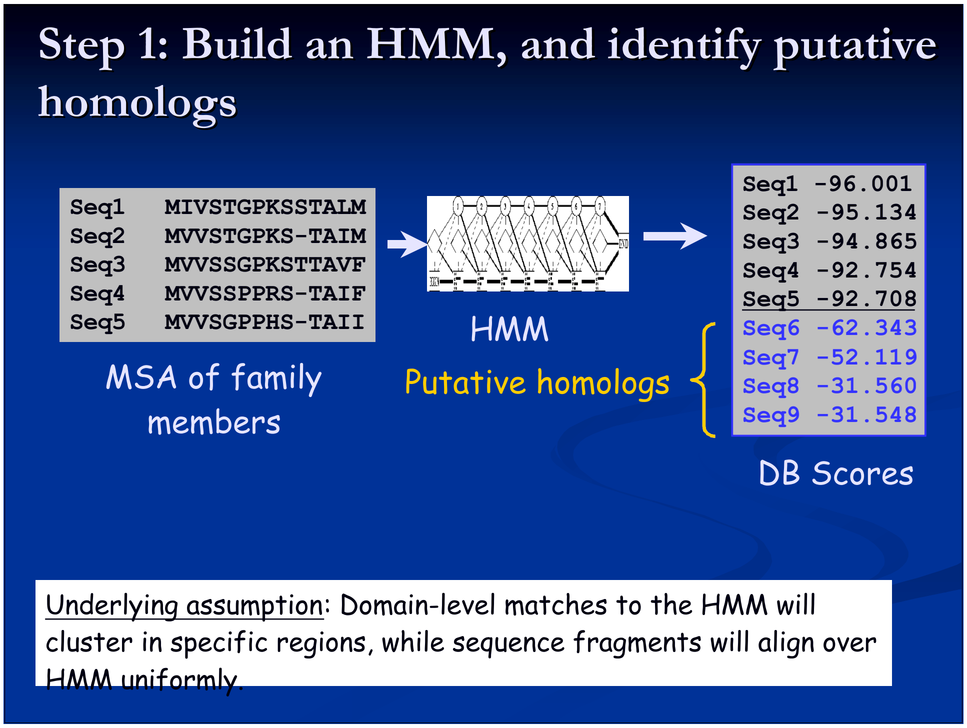
HMM performance improves with improved alignments (Preliminary results)





A Tale of Two Domains

Hidden Markov models,
Potassium channels,
Excursions in the Twilight Zone
and some stories about startups…



Step 1: Build an HMM, and identify putative homologs

Seq1	MIVSTGPKSSTALM
Seq2	MVVSTGPKS-TAIM
Seq3	MVVSSGPKSTTAVF
Seq4	MVVSSPPRS-TAIF
Seq5	MVVSGPPHS-TAII

MSA of family members

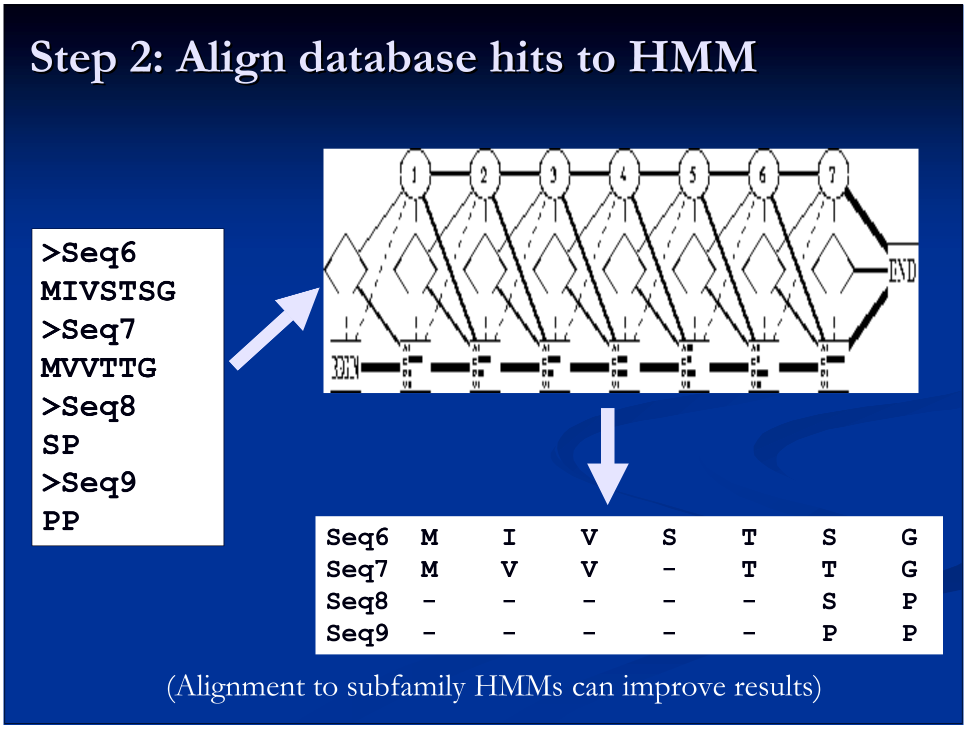
HMM

Seq1	-96.001
Seq2	-95.134
Seq3	-94.865
Seq4	-92.754
Seq5	-92.708
Seq6	-62.343
Seq7	-52.119
Seq8	-31.560
Seq9	-31.548

Putative homologs

DB Scores

Underlying assumption: Domain-level matches to the HMM will cluster in specific regions, while sequence fragments will align over HMM uniformly.

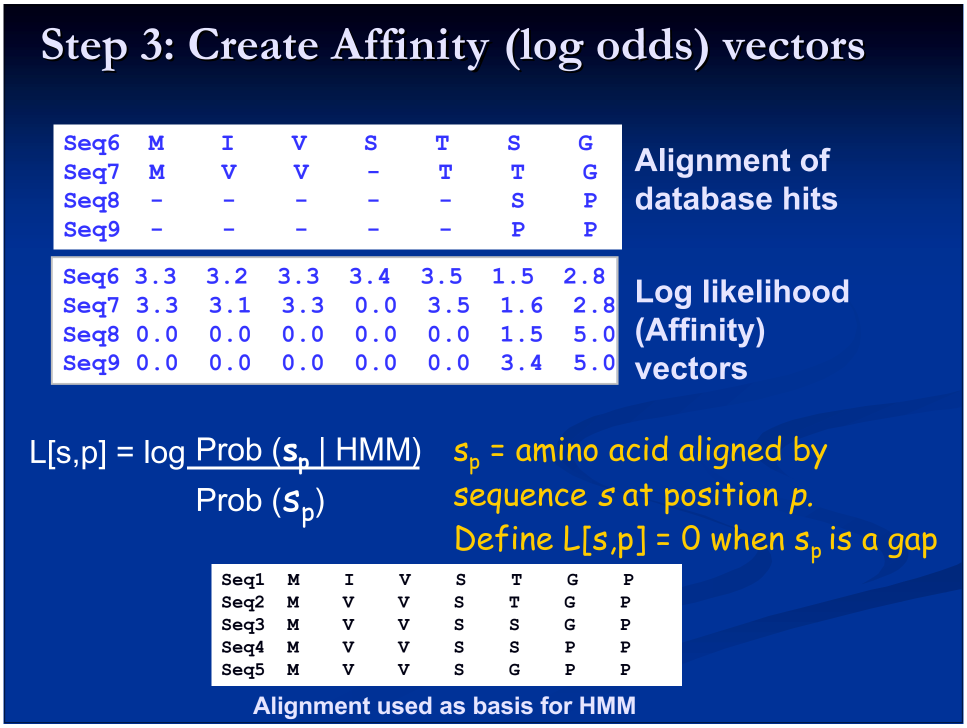


Step 2: Align database hits to HMM

```
>Seq6
MIVSTSG
>Seq7
MVVTTG
>Seq8
SP
>Seq9
PP
```

Seq6	M	I	V	S	T	S	G
Seq7	M	V	V	-	T	T	G
Seq8	-	-	-	-	-	S	P
Seq9	-	-	-	-	-	P	P

(Alignment to subfamily HMMs can improve results)



Step 3: Create Affinity (log odds) vectors

Seq6	M	I	V	S	T	S	G
Seq7	M	V	V	-	T	T	G
Seq8	-	-	-	-	-	S	P
Seq9	-	-	-	-	-	P	P

Alignment of database hits

Seq6	3.3	3.2	3.3	3.4	3.5	1.5	2.8
Seq7	3.3	3.1	3.3	0.0	3.5	1.6	2.8
Seq8	0.0	0.0	0.0	0.0	0.0	1.5	5.0
Seq9	0.0	0.0	0.0	0.0	0.0	3.4	5.0

Log likelihood (Affinity) vectors

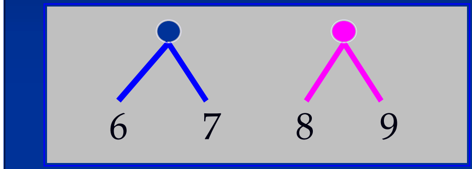
$$L[s,p] = \log \frac{\text{Prob } (s_p \mid \text{HMM})}{\text{Prob } (s_p)}$$

s_p = amino acid aligned by sequence s at position p.
Define $L[s,p] = 0$ when s_p is a gap

Seq1	M	I	V	S	T	G	P
Seq2	M	V	V	S	T	G	P
Seq3	M	V	V	S	S	G	P
Seq4	M	V	V	S	S	P	P
Seq5	M	V	V	S	G	P	P

Alignment used as basis for HMM

Step 4: Cluster affinity vectors

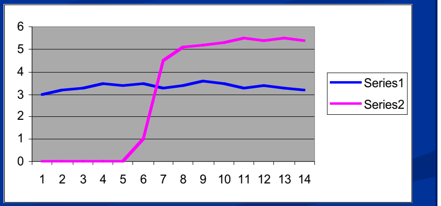


Clustering of vectors

Agglomerative clustering

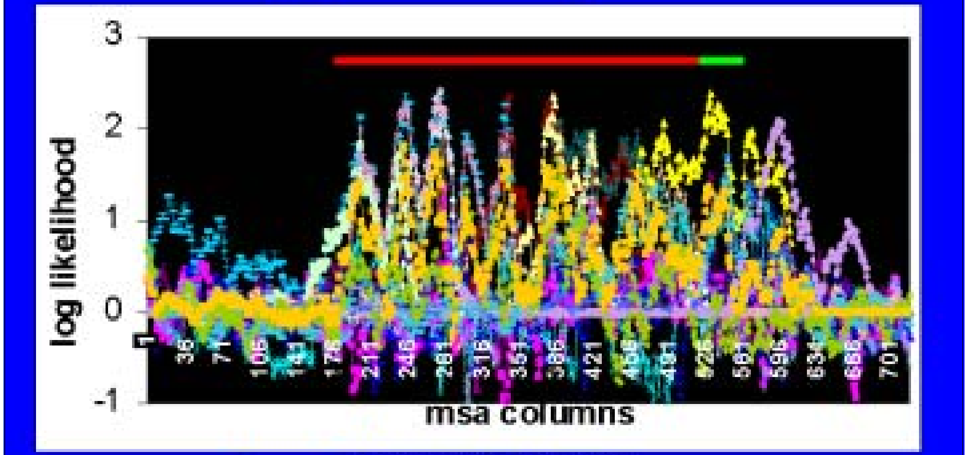
Euclidean distance

Cluster termination cutoff



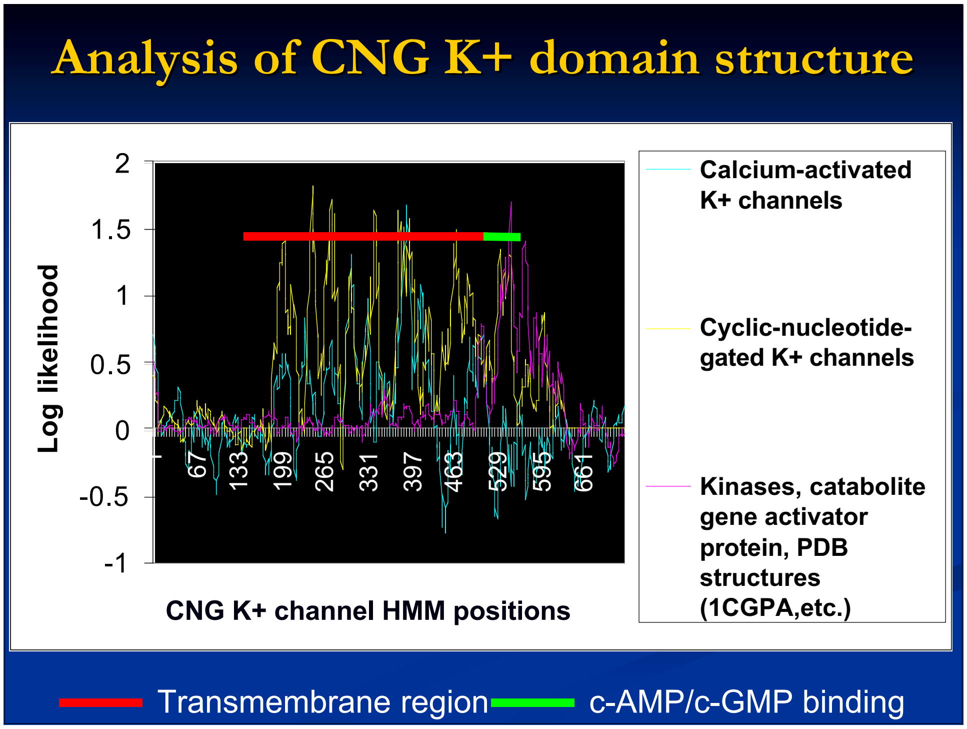
Plot LL clusters
(simple average)

Cyclic-Nucleotide-Gated K+ Channel HMM Alignment Analysis



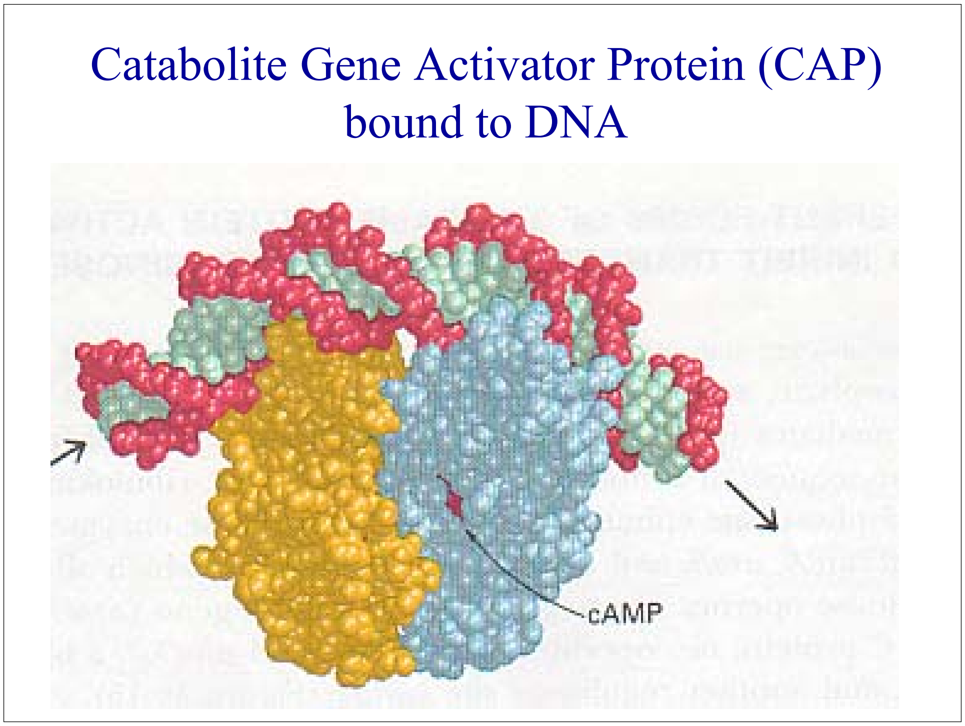
🟥 Transmembrane region
🟩 c-AMP/c-GMP binding

MAG p22

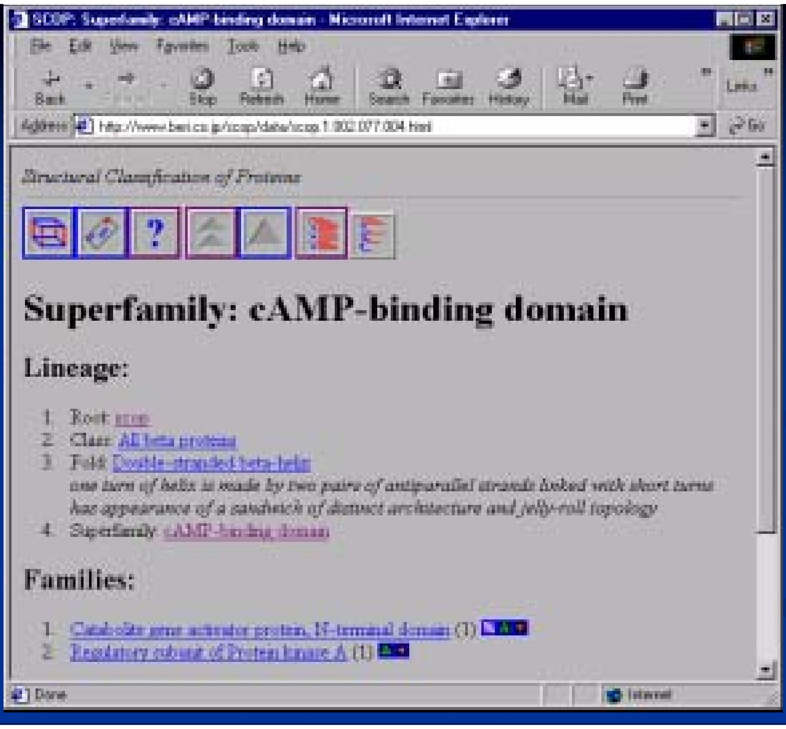


Analysis of CNG K+ domain structure

🟥 Transmembrane region 🟩 c-AMP/c-GMP binding



Catabolite Gene Activator Protein (CAP) bound to DNA



SCOP: Superfamily: cAMP binding domain - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Links
Address
Go
Done
Internet

http://www.bei.cs.jp/scop/data/scop.1.002.077.004.html

Structural Classification of Proteins

Superfamily: cAMP-binding domain

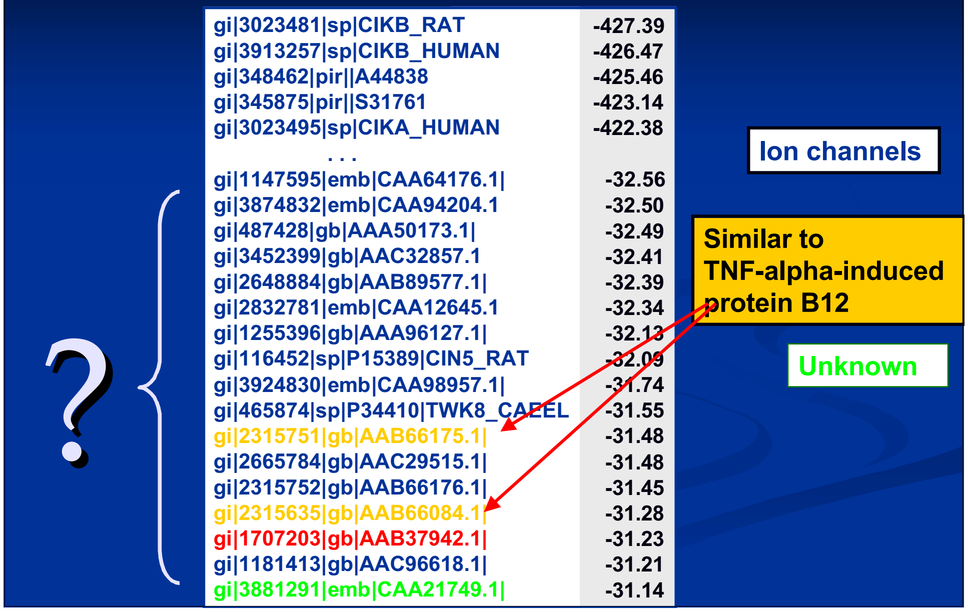
Lineage:

1. Root: scop
2. Class: All beta proteins
3. Fold: Double-stranded beta-helix
   one turn of helix is made by two pairs of antiparallel strands linked with short turns
   has appearance of a sandwich of distinct architecture and jelly-roll topology
4. Superfamily: cAMP-binding domain

Families:

1. Catabolite gene activator protein, N-terminal domain (1)
2. Regulatory subunit of Protein kinase A (1)

Database hits to Voltage-gated K+ channel



gi|3023481|sp|CIKB_RAT	-427.39
gi|3913257|sp|CIKB_HUMAN	-426.47
gi|348462|pir||A44838	-425.46
gi|345875|pir||S31761	-423.14
gi|3023495|sp|CIKA_HUMAN	-422.38
. . .	
gi|1147595|emb|CAA64176.1|	-32.56
gi|3874832|emb|CAA94204.1	-32.50
gi|487428|gb|AAA50173.1|	-32.49
gi|3452399|gb|AAC32857.1	-32.41
gi|2648884|gb|AAB89577.1|	-32.39
gi|2832781|emb|CAA12645.1	-32.34
gi|1255396|gb|AAA96127.1|	-32.13
gi|116452|sp|P15389|CIN5_RAT	-32.09
gi|3924830|emb|CAA98957.1|	-31.74
gi|465874|sp|P34410|TWK8_CAEEL	-31.55
gi|2315751|gb|AAB66175.1|	-31.48
gi|2665784|gb|AAC29515.1|	-31.48
gi|2315752|gb|AAB66176.1|	-31.45
gi|2315635|gb|AAB66084.1|	-31.28
gi|1707203|gb|AAB37942.1|	-31.23
gi|1181413|gb|AAC96618.1|	-31.21
gi|3881291|emb|CAA21749.1|	-31.14

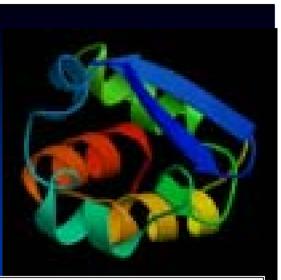
?

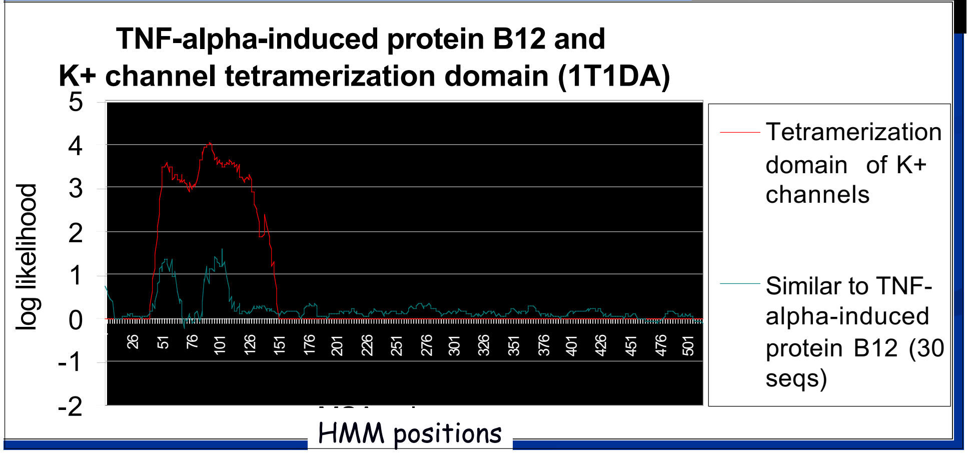
Ion channels

Similar to TNF-alpha-induced protein B12

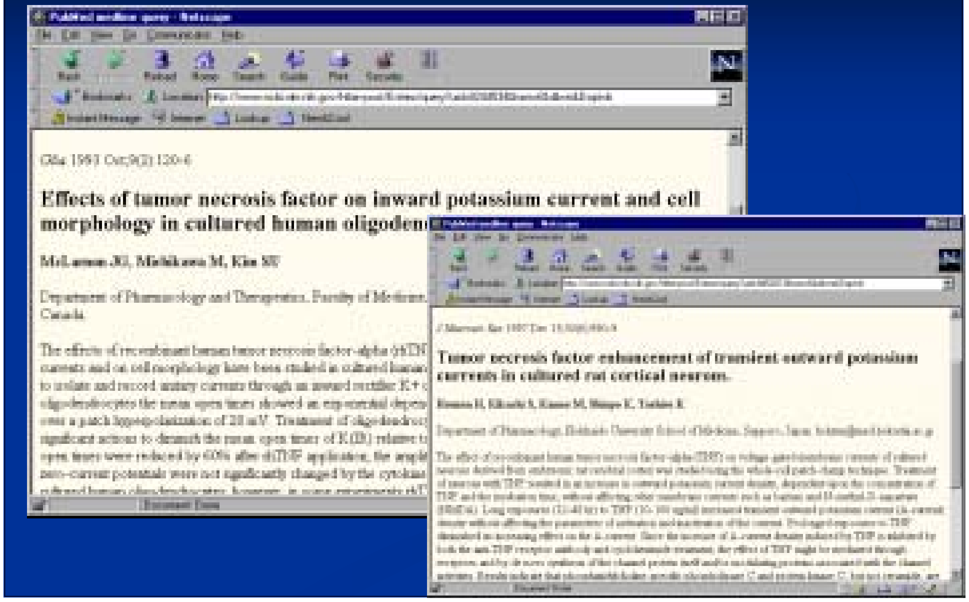
Unknown

Analysis of Voltage-gated K+ channels domain structure




TNF-alpha-induced protein B12 and
K+ channel tetramerization domain (1T1DA)
5
4
3
2
1
0
-1
-2
log likelihood
26
51
76
101
126
151
176
201
226
251
276
301
326
351
376
401
426
451
476
501
HMM positions
Tetramerization domain of K+ channels
Similar to TNF-alpha-induced protein B12 (30 seqs)

TNF-alpha acts on K+ current…but how?



Glia 1993 Oct;9(2):120-6

Effects of tumor necrosis factor on inward potassium current and cell morphology in cultured human oligodend

McLarnon JG, Michikawa M, Kim SU

Department of Pharmacology and Therapeutics, Faculty of Medicine, Canada.

The effects of recombinant human tumor necrosis factor-alpha (rhTN currents and on cell morphology have been studied in cultured human to isolate and record unitary currents through an inward rectifier K+ c oligodendrocytes the mean open times showed an exponential depen over a patch hyperpolarization of 20 mV. Treatment of oligodendroc significant actions to diminish the mean open times of K(IR) relative t open times were reduced by 60% after rhTNF application, the ampli zero-current potentials were not significantly changed by the cytokine [illegible]

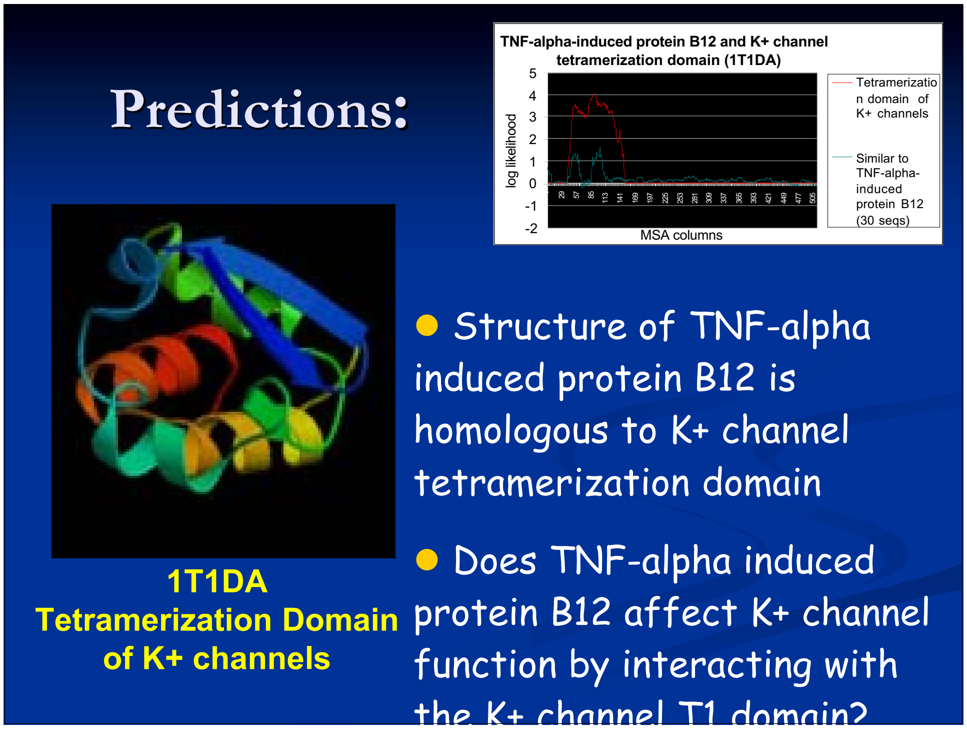
J Neurosci Res 1997 Dec 15;50(6):990-9

Tumor necrosis factor enhancement of transient outward potassium currents in cultured rat cortical neurons.

Houzen H, Kikuchi S, Kanno M, Shinpo K, Tashiro K

Department of Pharmacology, Hokkaido University School of Medicine, Sapporo, Japan. [illegible]

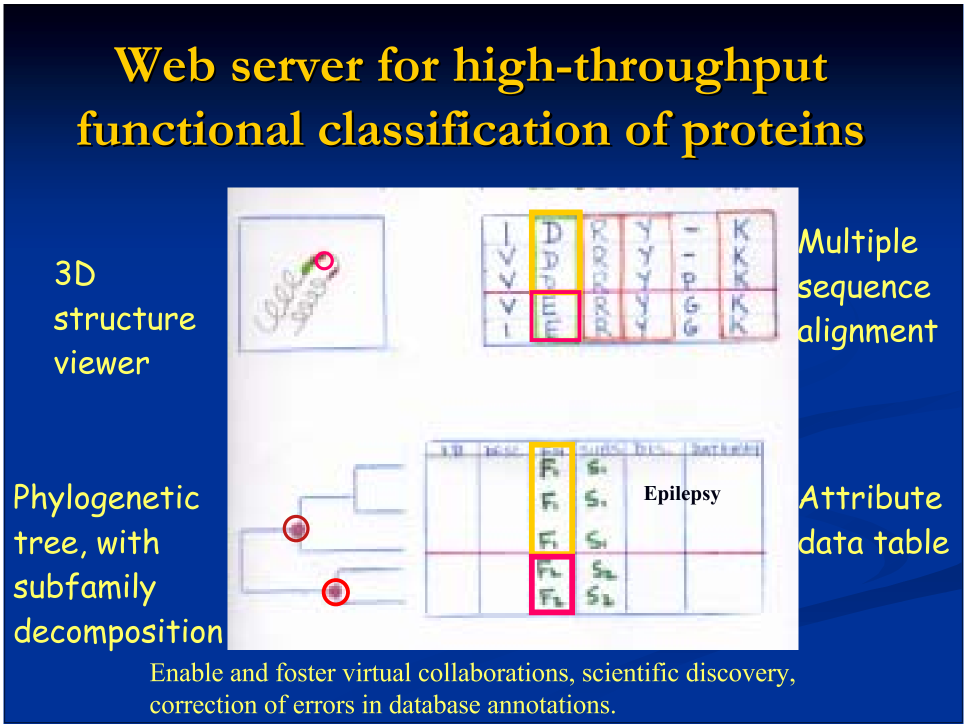
The effect of recombinant human tumor necrosis factor-alpha (TNF) on voltage-gated potassium currents of cultured neurons derived from embryonic rat cerebral cortex was studied using the whole-cell patch-clamp technique. Treatment of neurons with TNF resulted in an increase in outward potassium current density, dependent upon the concentration of TNF and the incubation time, without affecting other membrane currents [illegible]. Long exposure (12-48 hr) to TNF (10-100 ng/ml) increased transient outward potassium current (IA) current density without affecting the parameters of activation and inactivation of this current. Prolonged exposure to TNF [illegible] an increasing effect on the IA current. Since the increase of IA current density induced by TNF is inhibited by both the anti-TNF receptor antibody and cycloheximide treatment, the effect of TNF might be mediated through receptors and by de novo synthesis of the channel protein itself and/or modulating proteins associated with the channel activity. [illegible]



Predictions:

1T1DA
Tetramerization Domain of K+ channels

- Structure of TNF-alpha induced protein B12 is homologous to K+ channel tetramerization domain
- Does TNF-alpha induced protein B12 affect K+ channel function by interacting with the K+ channel T1 domain?



Web server for high-throughput functional classification of proteins

3D structure viewer

Multiple sequence alignment

Phylogenetic tree, with subfamily decomposition

Attribute data table

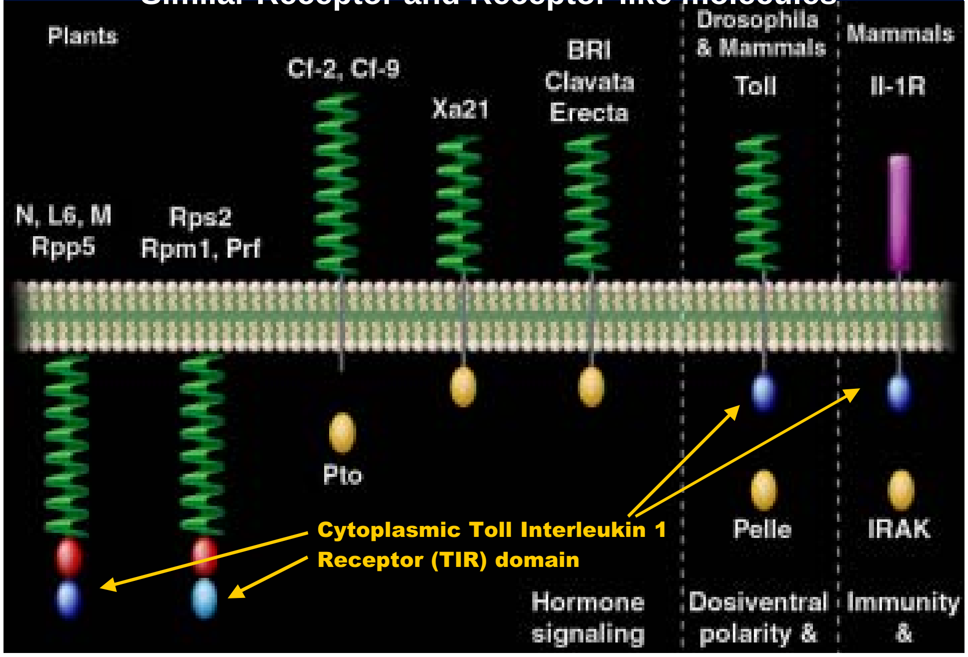
Enable and foster virtual collaborations, scientific discovery, correction of errors in database annotations.

Stalking disease-resistance proteins in rice

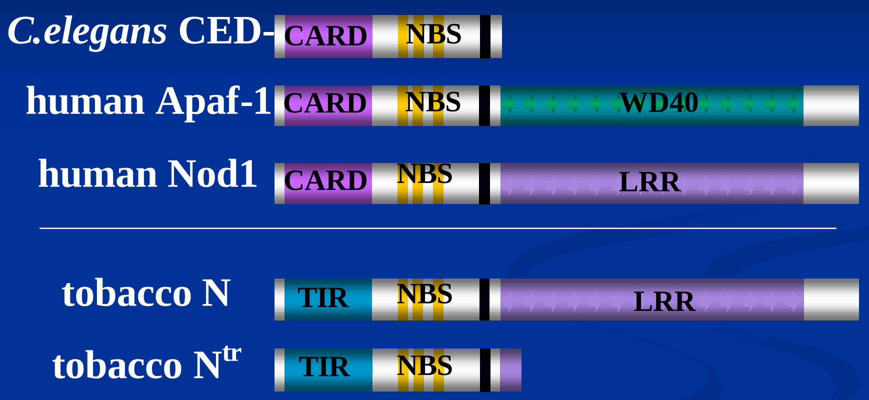
Toll Interleukin 1 domain
PDB structure 1FYX

Joint work with Barbara Baker and Brian Staskawicz

Plant and Animal Innate Immunity Mediated by Structurally Similar Receptor and Receptor-like molecules



Conserved “scaffolding” proteins in cell death



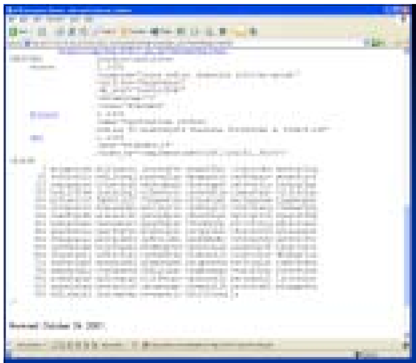
CARD: CAspase Recruitment Domain

Nod is implicated in CrohnÅs disease

N is involved in plant Hypersensitive Response (HR)

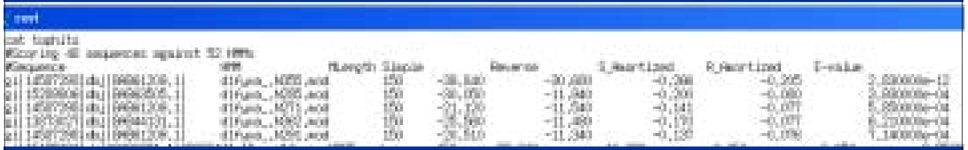
TIR domains missing from monocot species…

ÅMonocot sequences are absent from the TIR subfamilyÅ ÅToll/Interleukin-1 receptor homology (TIR), Å was entirely absent from monocot species in searches of both random genomic sequences and large collections of ESTsÅ



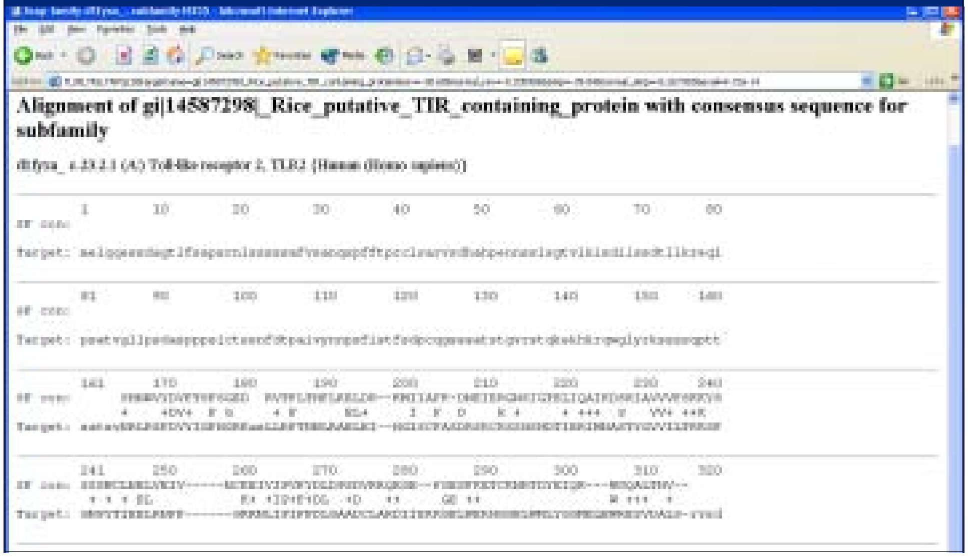
Searching the rice genome with general and subfamily HMMs for the TIR domain…

gi|14587298 gets top score (e-12 to TIR structure)



```
cat tophits
#Scoring 48 sequences against 52 HMMs
#Sequence                    HMM               MLength Simple   Reverse   S_Mortized  R_Mortized  E-value
gi|14587298|dbj|BAB61208.1|  d1fyxa_.N055.mod  150     -38.840  -30.680   -0.268      -0.205      3.83000e-12
gi|15238606|dbj|BAB62505.1|  d1fyxa_.N255.mod  150     -36.050  -11.940   -0.200      -0.080      3.60000e-04
gi|14587298|dbj|BAB61208.1|  d1fyxa_.N071.mod  150     -21.120  -11.540   -0.141      -0.077      5.85000e-04
gi|13975671|dbj|BAB44131.1|  d1fyxa_.N062.mod  150     -32.960  -11.480   -0.173      -0.077      6.22000e-04
gi|14587298|dbj|BAB61208.1|  d1fyxa_.N056.mod  150     -26.510  -11.340   -0.122      -0.076      7.14000e-04
```

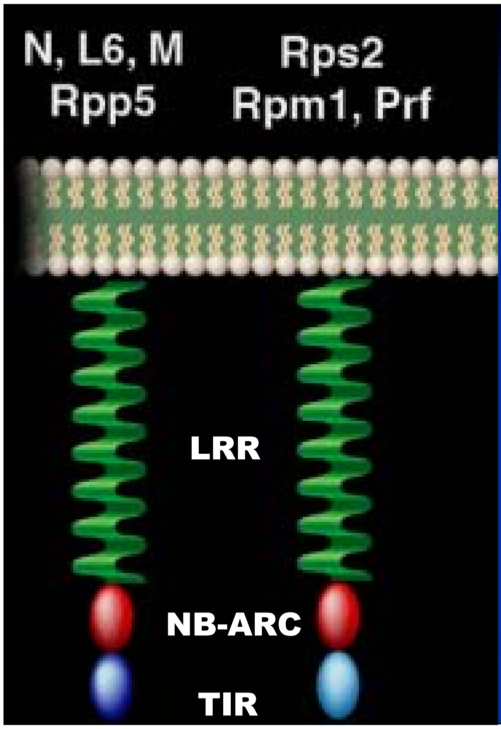
Alignment of rice sequence to Toll-like receptor 2 (TLR2) subfamily



Alignment of gi|14587298|_Rice_putative_TIR_containing_protein with consensus sequence for subfamily

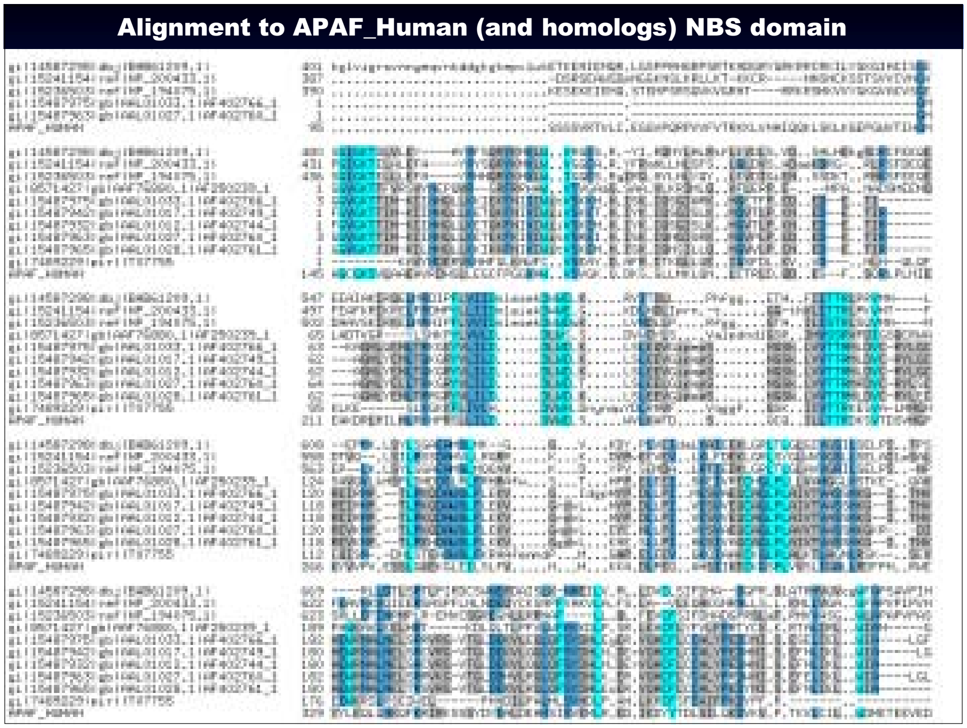
d1fyva_ c.23.2.1 (A) Toll-like receptor 2, TLR2 {Human (Homo sapiens)}

```
         1        10        20        30        40        50        60        70        80
SF cons:
Target:  [illegible]

         81       90        100       110       120       130       140       150       160
SF cons:
Target:  [illegible]

         161      170       180       190       200       210       220       230       240
SF cons: [illegible]
Target:  [illegible]

         241      250       260       270       280       290       300       310       320
SF cons: [illegible]
Target:  [illegible]
```



Where there's TIR there must be an NB-ARC domain

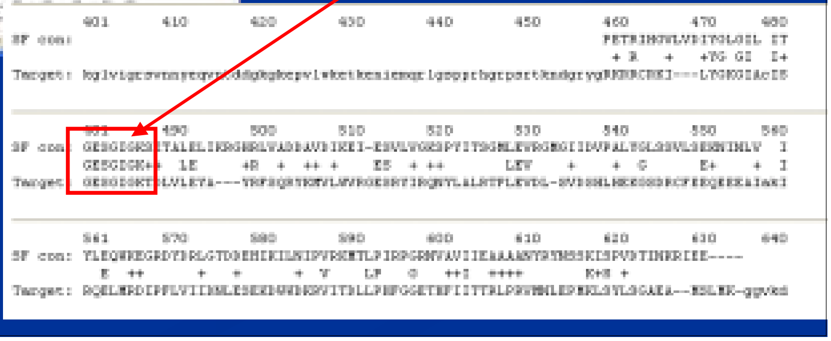
Alignment to APAF_Human (and homologs) NBS domain



Alignment to HPr kinase (serine kinase/phosphatase) with P-loop/Walker motif A

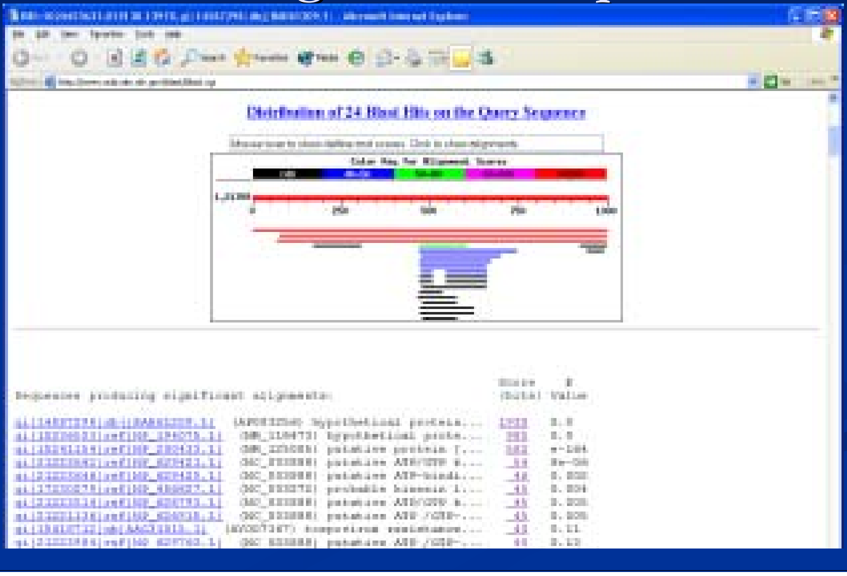
(possible structural similarity)

Phosphate binding motif



```
          401       410       420       430       440       450       460       470       480
SF con:                                                               FETRIHGVLVDITGLGIL IT
                                                                       + R    +    +TG GI  I+
Target:  kglvigrsvnnyeqvsddgkgkepvlwkeikenlemqrlgpgstrgrpsrtkndgrygRKRRCRKI---LTGKGIAcIE

          481       490       500       510       520       530       540       550       560
SF con:  GESGIGKSETALELIKRGHRLVADDAVEIKEI-ESVLVGKSPYITSGMLEVRGMGIIDVPALYGLSSVLSEKNIHLV  I
         GESGIGK++  LE      +R  +  ++ +     ES  + ++           LEV     +   +  G        E+     +  I
Target:  GESGIGKTHLVLEYA---YRFSQRYKMVLWVRGESRYIRQNYLALRTPLEVDL-SVISNLMEKGSDRCFERQEEKaIakI

          561       570       580       590       600       610       620       630       640
SF con:  YLEQWKEGRDYDRLGTDGEHIKILNIPVRKETLPIRPGRNVAVIIEAAAANYRYHSSKISPVDTINKRIEE----
            E  ++        +    +         +  V     LP    G     ++I   ++++               K+S  +
Target:  RQELWRDIPFLVIIKNLESEKDWWDKPVITRLLPRFGGETRFIITTRLPRWMNLEPRKLSYLSGAEA--ESLKK-ggvkd
```

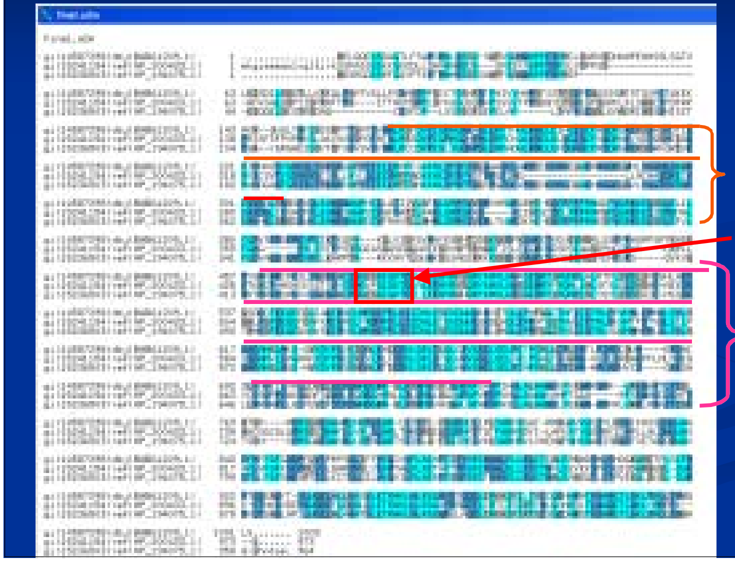
BLAST fails to detect TIR domain homologs for this sequence



Distribution of 24 Blast Hits on the Query Sequence

```
                                                          Score    E
Sequences producing significant alignments:               (bits) Value

[illegible]  [illegible] hypothetical protein...          [illegible]  0.0
[illegible]  [illegible] hypothetical prote...            [illegible]  0.0
[illegible]  [illegible] putative protein [...            [illegible]  e-104
[illegible]  [illegible] putative ATP/GTP b...            [illegible]  [illegible]
[illegible]  [illegible] putative ATP-bindi...            48           [illegible]
[illegible]  [illegible] probable kinesin l...            45           [illegible]
[illegible]  [illegible] putative ATP/GTP b...            45           [illegible]
[illegible]  [illegible] putative ATP /GTP-...            45           [illegible]
[illegible]  [illegible] [illegible] resistance...        43           0.11
[illegible]  [illegible] putative ATP /GTP-...            43           0.13
```

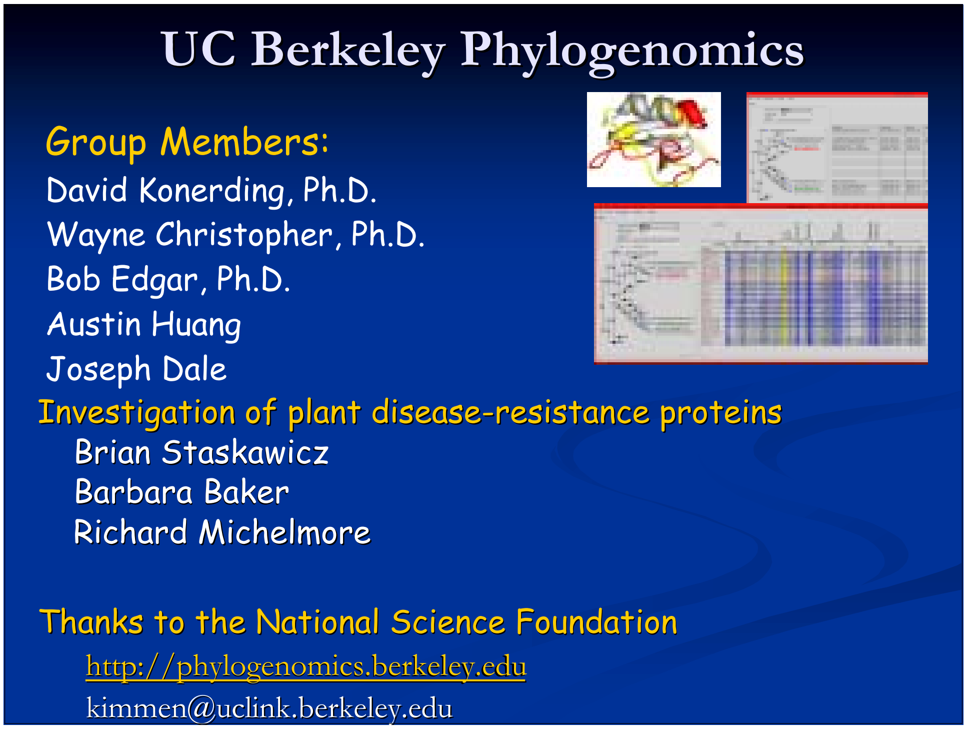
Clustering and aligning homologs with FlowerPower



TIR domain

Walker A motif

NBS-ARC domain



UC Berkeley Phylogenomics

Group Members:

David Konerding, Ph.D.

Wayne Christopher, Ph.D.

Bob Edgar, Ph.D.

Austin Huang

Joseph Dale

Investigation of plant disease-resistance proteins

Brian Staskawicz

Barbara Baker

Richard Michelmore

Thanks to the National Science Foundation

http://phylogenomics.berkeley.edu

kimmen@uclink.berkeley.edu